Statistical Inference for Optimal Transport

Tudor A. Manole

Department of Statistics and Data Science
Carnegie Mellon University
Pittsburgh, PA 15213

Thesis Committee
Sivaraman Balakrishnan (Co-Chair)
Larry Wasserman (Co-Chair)
Jing Lei
Dejan Slepčev
Axel Munk (Georg August University of Göttingen)
Jonathan Niles-Weed (New York University)
Alessandro Rinaldo (University of Texas, Austin)

Submitted in partial fulfillment of the requirements
for the degree of Doctor of Philosophy.

Keywords: Nonparametric Inference, Optimal Transport, Wasserstein Distance, Minimax Estimation, Semiparametric Efficiency, Density Estimation, High Energy Physics

For my family

Abstract

Optimal transport is a flexible framework for comparing probability distributions, which has received a recent surge of interest as a methodological tool in statistics. The aim of this thesis is to develop procedures for performing valid and efficient statistical inference for various objects arising from the optimal transport framework. On the one hand, we derive a semiparametric efficient estimator of the quadratic Wasserstein distance between probability measures of arbitrary fixed dimension. On the other hand, we develop a pointwise central limit theorem for the quadratic optimal transport map between multivariate periodic distributions. We also develop nonasymptotic and sequential inferential procedures for various optimal transport divergence functionals. These results provide a step toward the longstanding problem of performing practical inference for optimal transport in arbitrary dimension. Along the way, this thesis studies the related question of performing minimax estimation for optimal transport maps and costs, leading to new minimax upper or lower bounds for these problems in various settings. We close with an application of these ideas to a problem arising in experimental high energy physics, where we show that optimal transport can be used to address the problem of data-driven background estimation, arising in the search for new physical phenomena at the Large Hadron Collider.

Acknowledgments

I would like to begin by expressing my sincerest appreciation to my thesis advisors, Sivaraman Balakrishnan and Larry Wasserman. They have profoundly influenced my development as a researcher, my taste for statistical problems, and my form for scientific communication. They both bring a level of open-mindedness and lightheartedness to our work which makes research exciting and joyous. Siva has an uncanny ability to convey complicated concepts in the most intuitive of terms. Larry's encyclopedic knowledge and broad view of statistics is an inspiration. They never cease to impress me, and I consider myself extremely fortunate to have been their student. I can only hope to be able to pay forward some small fraction of what they have given to me. Thank you for everything, Siva and Larry.

I had the good fortune of collaborating closely with Jonathan Niles-Weed throughout most of my PhD. Jon is a consummate scholar, who has inspired me not only through his technical mastery, but also through the elegance with which he communicates scientific ideas, and the warmth he brings to every interaction. Thank you, Jon, for being a wonderful mentor, and for taking a chance on me all those years ago.

I am grateful to Axel Munk, Dejan Slepčev, Jing Lei, and Alessandro Rinaldo, for agreeing to be on my thesis committee, and for their kind advice and comments about my work. I would especially like to thank Axel, for very kindly hosting me in Göttingen on several occasions during my PhD. The developments in this thesis were motivated in large part by Axel's body of work on statistical optimal transport, and it is an honor to have the chance of collaborating with him. Above all else, Axel has inspired me through his constant inclination of staying grounded in the sciences, no matter how theoretical a project becomes.

I am very grateful to Aaditya Ramdas for collaborating with me on several projects throughout my PhD, and for being a constant source of advice and inspiration. Aaditya always pushed me to understand the big picture in every proof, every argument, every line—which is perhaps one of the most important skills for a doctoral student to learn.

I thank Mikael Kuusela, Patrick Bryant, and John Alison, for spurring my excitement for statistical applications in high energy physics. The enthusiasm they brought to our work was contagious. I have fond memories of our meetings, which often involved Patrick and John passionately writing complicated Feynman diagrams on Larry's blackboard, while Larry and I slowly tried to keep up, with the help of Mikael's translation.

I am very grateful to Nhat Ho for our collaboration during the early years of my PhD. His papers on finite mixture models were among the first that I ever read as an undergraduate, and it was my great pleasure to build upon these ideas in our joint work.

Abbas Khalili has been a wonderful mentor ever since my undergraduate years at McGill. He introduced me to the field of statistics, showed me its beauty, and was ultimately responsible for my choosing to pursue a PhD in this field. He always went above and beyond in making time to train me, meticulously teaching me to simulate, derive, write, and revise—always to perfection. I will always be influenced by the formative years I spent working with him, and I cannot thank him enough for his guidance and friendship.

I would like to thank two good friends in the optimal transport community: Aram Pooladian and Shayan Hundrieser. My friendship with Aram dates back to our undergraduate years at McGill, and to this day he is always available to get on a call with me, and even to check my proofs if I ask nicely. My friendship with Shayan developed over the course of visits to each others' departments in the last two years of our PhDs. I have fond memories from our travels and long research discussions, and look forward to many more.

I would like to thank all the faculty and students in the CMU Department of Statistics and Data Science, who make it such a warm, open-minded, and exciting place to do research. I want to specifically thank Addison Hu, Ian Waudby-Smith, and my academic brothers Beomjo Park and Lucas Kania, for all the wonderful memories we shared. Pursuing a PhD could not be more joyous when surrounded by friends like yourselves. I would also like to thank Neil Spencer, Niccolò Dalmasso, and Kevin Lin for their kind advice over the years. There are many other students in the department whom I want to thank, among which are: Maya Shen, Kenta Takatsu, Siddhaarth Sarkar, Alec McClean, Sasha Podkopaev, Nil-Jana Akpinar, Mikaela Meyer, Kayla Scharfstein, Selina Carter, Julia Walchessen, Catherine Wang, and Kyle Schindl.

Thank you to my three closest friends—Daniel Audi, Alexandre Mizon, and Makoto Rheault-Kihara—for all the happy memories we shared, and for always being there to pick up right where we left off, even when I was away for years on end.

I am grateful to Renaud Schweizer for cultivating my early interest in mathematics. The world would be inundated with mathematicians if the subject were always taught in Renaud's style.

My heartfelt thanks go to my parents, Liana and Silvan, and to my sister Eva, for their unwavering love and patient support. All of my achievements are ultimately to their credit. My parents made many sacrifices in order to give me opportunities which they did not have during their youth, and it is only thanks to their efforts that I have the chance of pursuing an academic career. For this, I will always be grateful.

And finally, Megan, words cannot begin to express my deepest gratitude for our wonderful years together. You will always have my endless love and admiration.

Contents

List of Figures

Chapter 1

Introduction

The field of optimal transport arose from a resource allocation problem first posed by Monge (1781), and has since developed into a rich branch of mathematical analysis (Ambrosio, Gigli, and Savare, 2008; Villani, 2008; Figalli and Glaudo, 2021), shaped by important historical influences from economics (Kantorovich, 1942; Galichon, 2016) and physics (Jordan, Kinderlehrer, and Otto, 1998; Benamou and Brenier, 2000).

In statistical contexts, optimal transport has historically been used as a theoretical tool for asymptotic theory (Bickel and Freedman, 1981; Shao and Tu, 2012), since it gives rise to a convenient characterization of weak convergence over the space of probability measures. In recent years, however, optimal transport has proven to be more than just a useful theoretical device, and has gained widespread popularity as a *methodological* tool in statistics. Such methodological applications date back at least to the early paper of Munk and Czado (1998), who proposed the use of optimal transport to solve a univariate bioequivalence hypothesis testing problem. This work was followed up by several other related papers on univariate nonparametric testing (e.g. Freitag, Czado, and Munk (2007) and references therein), but it was not until the past five to ten years that the optimal transport framework began to receive widespread adoption in the statistics community. This surge of recent interest was driven in part by computational advances (Cuturi, 2013; Peyré and Cuturi, 2019), the availability of inferential procedures beyond univariate measures (Sommerfeld and Munk, 2018), and successful applications in the machine learning and computer vision communities (e.g. Arjovsky, Chintala, and Bottou (2017)). Optimal transport has now been used as a methodological tool for a variety of statistical tasks, ranging from multivariate distribution-free testing, to inference over the space of probability measures. We refer to Kolouri et al. (2017), Panaretos and Zemel (2019a, 2019b), Hallin (2022), for surveys of recent developments.

In order to see why optimal transport has been such a successful data analytic tool, let us provide a heuristic introduction to the subject, deferring a more formal description to Section 1.3 below. In its most basic formulation, the optimal transport problem is a question about *transport maps* between probability measures. Concretely, given two absolutely continuous probability distributions P and Q supported in $\mathbb{R}^d$, a transport map from P to Q is a vector field T which satisfies the relation

$$T(X) \sim Q, \quad \text{for any random variable } X \sim P. \tag{1.1}$$

Thus, a transport map T is simply a transformation of a random variable $X \sim P$ onto a second random variable $Y = T(X)$ whose distribution is Q. Said differently, Q is a pushforward of P under the map T.

A wide range of statistical applications involve transforming random variables to ensure they follow a desired distribution, and transport maps play a natural role in such applications. As a classical example, the technique of quantile-quantile (Q-Q) plots consists of assessing the discrepancy between two univariate distributions by visualizaing a transport map between them, and determining the extent to which it deviates from the identity. Another example is the probability integral transform, which is a procedure for sampling from univariate probability measures, based on the fact that the quantile function of a univariate measure Q is a transport map from the uniform distribution $P = \mathcal{U}[0, 1]$ onto Q. More generally, a popular approach in machine learning for high-dimensional sampling is to construct a suitable transport map between a given reference distribution P, and the distribution of interest, say Q. To sample from the complex measure Q, it then suffices to draw observations from the simple measure P, and push them through the map T. In general, there may exist infinitely-many transport maps T that one can choose to carry out this procedure, and different choices have led to extremely successful sampling procedures, such as the methods of normalizing flows (Kobyzev, Prince, and Brubaker, 2020) or denoising diffusion probabilistic models (Ho, Jain, and Abbeel, 2020).

One of the classical results in the optimal transport framework is that, among the potentially infinite collection of transport maps between two given distributions, there exists one which is most parsimonious (in a suitable sense), and which satisfies a multivariate notion of monotonicity. This object is called the *Brenier map*, or the *(quadratic) optimal transport map*, and can be viewed as a default choice of transport map for applications where no other canonical choice is available. There are several equivalent ways of defining the Brenier map; the following is one approach, that arises from the celebrated paper of Brenier (1991).

Theorem 1 (Brenier's Polar Factorization Theorem (Informal)). *For any absolutely continuous probability measures P and Q on $\mathbb{R}^d$, there exists a unique gradient of a convex function $T_0 = \nabla\varphi_0$ which pushes forward P onto Q. Furthermore, for any transport map T from P to Q, the following decomposition holds*

$$T = T_0 \circ S,$$

where S is a measure-preserving map, i.e. a transport map from P onto itself. T_0 is called the Brenier map, or (quadratic) optimal transport map, from P to Q.

By analogy to univariate convex functions, whose derivatives are always monotonic (when they exist), it is natural to think of gradients of multivariate convex functions as satisfying a multivariate notion of monotonicity. In this language, Brenier's polar factorization theorem implies that for any given measures P and Q, there exists a unique monotonic vector field T_0 which pushes P forward onto Q. Furthermore, the theorem shows that any other transport map T from P to Q can be written as a composition of this monotonic vector field T_0, and a map S which is redundant as far as the problem of transporting P to Q is concerned. It is in this sense that T_0 is most parsimonious among all transport maps, since it is the one for which the redundant component is $S = \mathrm{Id}$.

Later on, we will see that the Brenier map can also be characterized in terms of an optimization problem: it can be shown that T_0 is the unique solution to the so-called *Monge optimal transport problem*, defined by

$$\mathcal{T}_c(P,Q) = \min_{T:\mathbb{R}^d \to \mathbb{R}^d} \mathbb{E}\big[c(X, T(X))\big], \quad \text{subject to } \ T(X) \sim Q, \tag{1.2}$$

where $c(x,y) = \|x-y\|^2$ for all $x, y \in \mathbb{R}^d$, and $X \sim P$. Equation (1.2) characterizes the Brenier map as the $L^2(P)$ projection of the identity onto the space of transport maps from P to Q. This provides a different sense in which T_0 is most parsimonious. One could also consider different types of projections by choosing a different *cost function* $c : \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}_+$ in the above display, but this leads to different notions of optimal transport maps which do not enjoy the properties set forth in Theorem 1.

Let us begin by illustrating the statistical utility of Brenier maps, by showing how they can be used to address the longstanding problem of defining multivariate analogues of traditional notions from univariate statistics, such as quantiles functions, cumulative distribution functions, and their empirical analogues, based on ranks and signs. As mentioned previously, it is well-known that the quantile function G^{-1} of a univariate probability measure Q is a transport map from the uniform distribution $P = \mathcal{U}[0,1]$ onto Q. Recalling that quantile functions are monotonic, we deduce from Theorem 1 that G^{-1} must, in fact, be the unique Brenier map which pushes forward P onto Q. This suggests a very simple definition for a multivariate quantile: given a measure Q supported in $\mathbb{R}^d$, and given a reference measure P, such as the uniform distribution on $[0,1]^d$, it is natural to *define* the quantile function of Q as the unique optimal transport map pushing forward P onto Q. This map is referred to as the *Monge-Kantorovich quantile function* of Q. Notions of cumulative distribution functions, ranks, and signs, can then be defined using a similar principle. These definitions were first proposed by Chernozhukov et al. (2017); Hallin et al. (2021), who argue that they retain essentially all of the desirable properties of their univariate counterparts which makes them useful tools for distribution-free inference. A great deal of follow-up work has then used these definitions to propose multivariate rank-based, distribution-free procedures for testing hypotheses of goodness-of-fit, two-sample equality, independence, and symmetry, which often retain the same asymptotic efficiency properties as their univariate counterparts (cf. Shi, Drton, and Han (2020); Deb, Ghosal, and Sen (2021); Deb, Bhattacharya, and Sen (2021); Huang and Sen (2023)). We refer to Hallin (2022) for a survey. Beyond these examples, optimal transport maps have been used in a variety of other recent methodological tasks, such as domain adaptation (Courty et al., 2016; Redko et al., 2019; Rakotomamonjy et al., 2022; Zhu et al., 2021), distributional regression (Ghodrati and Panaretos, 2022), and generative modeling (Finlay et al., 2020; Onken et al., 2021).

Although we have primarily focused on optimal transport maps up to this point, an equally important quantity for statistical applications turns out to be the optimal objective value of the optimization problem (1.2). The quantity $\mathcal{T}_c(P, Q)$ has the natural interpretation of measuring the expected cost in transporting probability mass from P forward onto Q, and thus defines an interpretable notion of divergence between these measures. In the special case where the cost function c is of the form $c(x, y) = \|x - y\|^r$, for some $r \geq 1$, we write

$$W_r := \left(\mathcal{T}_{c_r}\right)^{1/r},$$

and the functional W_r turns out to define a metric over the space of probability measures on $\mathbb{R}^d$ with finite moments up to order r, known as the r-Wasserstein distance. In its most general formulation that we will give in Section 1.3 below, the Wasserstein distance does not presume distributions which are absolutely continuous with respect to a common dominating measure. Furthermore, contrary to other classical metric between probability measures, such as the Hellinger or Total Variation distances, the Wasserstein distance is sensitive to the underlying geometry of the measures being compared, due to the Euclidean norm embedded into its definition. These considerations make it a natural and powerful data analytic tool. For example, the empirical Wasserstein distance has been advocated as a natural test statistic for goodness-of-fit or two-sample testing Munk and Czado (1998); Ramdas, Trillos, and Cuturi (2017); Sommerfeld and Munk (2018); Hallin, Mordant, and Segers (2021). Wasserstein distances are also a natural tool for metrizing the space of mixing measures in latent variable models (Nguyen, 2013; Ho, Yang, and Jordan, 2022; Wang, 2019), an observation which has both been used for the theoretical purpose of deriving convergence rates of estimation procedures (cf. Chen (2023) for a survey), and for the methodological purpose of defining new model selection methods (cf. Do et al. (2024) and references therein). Let us also mention that Wasserstein distances have been used for a variety of other tasks, such as distributional clustering (Ho et al., 2017; Verdinelli and Wasserman, 2019), performing inference from simulations (Bernton et al., 2019), minimum distance estimation (Bernton et al., 2019; Zhang and Chen, 2022), analyzing exchangeable random graphs (Lei, 2021), and performing inference over the space of probability measures (Petersen and Müller, 2019; Panaretos and Zemel, 2019).

Finally, we emphasize that there is an increasing number of statistical applications where the optimal transport problem has a scientific interpretation, and thus becomes the object of interest in its own right. For instance, the Wasserstein distance has recently been used in experimental high energy to quantify the distance between histograms of energy deposits produced by collider events (Komiske et al., 2019), while optimal transport maps have been used in biological applications to model the development of embryonic cells (Schiebinger et al., 2019), and in biophysics to quantify the colocalization of cellular samples (Tameling et al., 2021). In these applications, optimal transport quantities become the primary target of statistical inference.

1.1 What this Thesis is About

The preceding applications raise the important question of performing statistical inference for objects in the optimal transport framework. Concretely, if a practitioner has access to i.i.d. samples

$$X_1, \dots, X_n \sim P, \quad Y_1, \dots, Y_m \sim Q \tag{1.3}$$

from unknown measures P and Q, how can they perform statistical inference for objects like Wasserstein distances and optimal transport maps? For example, if a statistician wishes to report a point estimate of a Monge-Kantorovich median, how can they also provide a confidence interval centered at their estimate? If a biophysist wishes to use optimal transport to quantify the colocalization of two protein samples, how can they test whether the level of colocalization is nonzero? How does one optimally test hypotheses separated in Wasserstein distance? This thesis is broadly motivated by inferential questions of this type.

> **Thesis Statement:** To develop procedures for performing valid and efficient statistical inference for optimal transport maps and optimal transport costs, between probability measures of arbitrary fixed dimension.

The subject of inference for optimal transport was still in its infancy when this thesis began, but has since been the subject of intensive study in the literature. Let us briefly highlight some of these recent developments, deferring a more thorough survey to the individual chapters which will follow.

Perhaps the object which has received the most attention in recent literature has been the *empirical Wasserstein distance*, which is the natural sample analogue of the Wasserstein distance. As mentioned previously, the study of limiting distributions for the empirical Wasserstein distance was initiated in the univariate setting by Munk and Czado (1998), and followed up by the works of del Barrio et al. (1999); Freitag, Munk, and Vogt (2003); Samworth and Johnson (2005); del Barrio, Giné, and Utzet (2005); Freitag and Munk (2005); Freitag, Czado, and Munk (2007), as well as the more recent works of del Barrio, Gordaliza, and Loubes (2019); Berthet, Fort, and Klein (2020). Each of these papers heavily exploit the one-dimensional structure of the Wasserstein distance, which allows it to be expressed as the L^r distance between the quantile functions of the distributions to be compared. This allows for limit laws for the empirical Wasserstein distance to be derived using the classical asymptotic theory of the empirical quantile process.

The first studies of the empirical Wasserstein distance beyond the univariate setting were the works of Sommerfeld and Munk (2018) and Tameling, Sommerfeld, and Munk (2019), which considered the case where the measures are supported on a finite or countably infinite metric space. This setting already provides a significant departure from that of univariate measures, requiring new insights to deal with the lack of total ordering of the underlying space, and the resulting lack of Hadamard smoothness of the Wasserstein functional. The ideas in these works have also been extended to the situation where at least one of the two measures P and Q is finitely-supported (Sadhu, Goldfeld, and Kato, 2023; del Barrio, González-Sanz, and Loubes, 2024).

Moving now to the situation where both P and Q are genuinely d-dimensional distributions, say absolutely continuous with respect to the Lebesgue measure, a fundamental contribution was made by del Barrio and Loubes (2019), who showed that, under mild regularity conditions, the quadratic empirical Wasserstein distance $W_2^2(P_n, Q_m)$, between independent empirical measures P_n and Q_m comprised of the i.i.d. samples (1.3), enjoys a central limit theorem centered at its expectation in arbitrary dimension $d \geq 1$:

$$\sqrt{\frac{nm}{n+m}}\big(W_2^2(P_n, Q_m) - \mathbb{E}W_2^2(P_n, Q_m)\big) \rightsquigarrow N(0, \sigma^2), \tag{1.4}$$

for some $\sigma^2 \geq 0$, which is positive if and only if $P \neq Q$. This result was also extended to more general optimal transport costs by del Barrio, González-Sanz, and Loubes (2021). Although these results are very useful, their centering sequence is a barrier to their use for inferential purposes. Indeed, we will see in Chapter 3 below that the bias of the empirical Wasserstein distance in general dimension typically satisfies

$$\mathbb{E}W_2^2(P_n, Q_m) - W_2^2(P, Q) \gtrsim (n \wedge m)^{-2/d}.$$

This rate is of leading order whenever $d > 4$, and in this case, the centering sequence in equation (1.4) cannot be replaced by the population Wasserstein distance. This precludes the possibility of constructing an asymptotic confidence interval using equation (1.4) when $d > 4$. Nevertheless, we emphasize that limit laws with desirable centering are available in the low-dimensional regime $d \leq 3$, and more generally when only one of the two measures P and Q is of low intrinsic dimension (Hundrieser et al., 2022). One of the contributions of this thesis will be to show, in Chapter 7 below, that there exists a distinct estimator of the Wasserstein distance which enjoys a central limit theorem centered at its population counterpart in any dimension $d \geq 1$, under appropriate smoothness and support assumptions. To the best of our knowledge, this leads to the first procedure for constructing confidence sets for the Wasserstein distance in general dimension.

Beyond the Wasserstein distance, let us emphasize that the question of inference for other optimal transport divergence functionals—such as the Sinkhorn divergence, the Sliced Wasserstein distance, and the Gaussian-Smoothed Wasserstein distance—turns out to be significantly more tractable than that of the vanilla Wasserstein distance, and has received a comprehensive treatment in the literature—see for instance Mena and Weed (2019), Goldfeld and Greenewald (2020), Sadhu, Goldfeld, and Kato (2021), González-Sanz, Loubes, and Niles-Weed (2022), del Barrio et al. (2023), González-Sanz and Hundrieser (2023), and Goldfeld et al. (2024a, 2024b, 2024c).

The question of performing inference for optimal transport maps has received significantly less attention. In the one-dimensional case, uniform confidence bands for optimal transport maps can be derived using classical strong approximation theory for the quantile-quantile process (Aly, 1986), or using asymptotics for kernel-based estimators (Ponnoprat, Okano, and Imaizumi, 2024). In the case where P and Q are discrete distributions, Klatt, Munk, and Zemel (2022) have derived central limit theorems for optimal transport couplings, while del Barrio, González-Sanz, and Loubes (2024) and Sadhu, Goldfeld, and Kato (2023) considered the case where only one of P or Q is discrete. Let us also emphasize that uniform limit laws for regularized variants of the Brenier map are well-studied—see for instance Harchaoui, Liu, and Pal (2020), Gunsilius and Xu (2021), González-Sanz, Loubes, and Niles-Weed (2022), Goldfeld et al. (2024), and González-Sanz and Hundrieser (2023). Nevertheless, we are not aware of any prior work on the question of performing inference for Brenier maps themselves, when the underlying measures are absolutely continuous and of dimension greater than one. Our work in Chapter 6 provides a first step in this direction, by deriving pointwise limit laws for an estimator of Brenier maps on the d-dimensional flat torus, when $d \geq 3$.

1.2 Organization of this Thesis

Let us now provide a chapter-by-chapter summary of the various contributions of this thesis.

Chapter 2: Minimax Confidence Intervals for the Sliced Wasserstein Distance.

In this chapter, we take a preliminary step toward the question of performing statistical inference for the Wasserstein distance, by studying a related functional known as the Sliced Wasserstein distance. Introduced by Rabin et al. (2011); Bonneel et al. (2015), this metric is obtained by averaging the Wasserstein distance between random one-dimensional projections of the distributions to be compared. The resulting metric shares some of the same qualitative behaviour as the multivariate Wasserstein distance, but is significantly easier to compute, since the Wasserstein distance between univariate projections is available in closed-form. We will see that slicing also has statistical advantages: Our results imply that the minimax rate of estimating the sliced distance is significantly faster than that of estimating the multivariate Wasserstein distance itself. We will also show that the definition of the Sliced Wasserstein distance makes it easily amenable to performing inference. Indeed, on the one hand, we construct confidence intervals for the sliced distance which have finite-sample validity under very mild assumptions, based on self-normalized concentration inequalities for the empirical quantile process. On the other hand, under somewhat stronger regularity conditions, we will show that the sample analogue of the Sliced Wasserstein distance enjoys a $\sqrt{n}$-central limit theorem centered at its population counterpart, and we establish the validity of the bootstrap in estimating this limiting distribution. Altogether, these results provide a relatively comprehensive set of tools for performing inference for the Sliced Wasserstein distance.

The contents of this chapter are adapted from the following publication:

- Manole, T., Balakrishnan, S., and Wasserman, L. (2022). Minimax Confidence Intervals for the Sliced Wasserstein Distance. *Electronic Journal of Statistics*, 16:2252–2345.

Chapter 3: Sharp Rates for Empirical Optimal Transport with Smooth Costs.

In this chapter, we return our attention to the multivariate Wasserstein distance, and more generally, to a broad class of optimal transport functionals $\mathcal{T}_c$ with respect to a convex ground cost function c. Our aim is to derive sharp upper and lower bounds on the minimax risk of estimating $\mathcal{T}_c(P, Q)$ for probability measures P, Q which are of arbitrary (fixed) dimension. To avoid several corner cases arising in low dimension, we focus on the regime $d \geq 5$.

It is well known that an empirical measure comprising n independent observations from an absolutely continuous distribution on $\mathbb{R}^d$ converges to that distribution at the rate $n^{-1/d}$ in

Wasserstein distance, which can be used to prove that the sample analogue of the functional $\mathcal{T}_c$ converges at this same rate, for a wide range of cost functions c. However, we show somewhat surprisingly that when the cost c is smooth, this analysis is loose: plug-in estimators based on empirical measures may converge as much as quadratically faster, at the rate $n^{-2/d}$. This finding generalizes a result for the quadratic optimal transport cost proven by Chizat et al. (2020) concurrently to this thesis. As a corollary, we show that the Wasserstein distance between two distributions is significantly easier to estimate when the measures are well-separated. Building upon work of Niles-Weed and Rigollet (2022), we also prove corresponding minimax lower bounds, which altogether lead to the following characterization of the minimax functional estimation risk:

$$(n \log n)^{-\alpha/d} \lesssim \inf_{\widehat{\mathcal{T}}_n} \sup_{(P,Q)} \mathbb{E}_{(P,Q)} \big|\widehat{\mathcal{T}}_n - \mathcal{T}_c(P,Q)\big| \lesssim n^{-\alpha/d} \quad (d \geq 5), \tag{1.5}$$

where the cost function c is assumed to satisfy several structural properties, including the smoothness assumption $c \in \mathcal{C}^\alpha$ for some $\alpha \in (0, 2]$. Furthermore, the infimum in the above display is over all estimators of $\mathcal{T}_c(P,Q)$ based on independent i.i.d. samples $X_1, \dots, X_n \sim P$ and $Y_1, \dots, Y_n \sim Q$, and the supremum is over all pairs of probability measures supported in a fixed convex and compact subset of $\mathbb{R}^d$. Our proofs rely on empirical process theory arguments based on tight control of L^2 covering numbers for locally Lipschitz and semi-concave functions. As a byproduct of our proofs, we derive L^∞ estimates on the displacement induced by the optimal coupling between any two measures satisfying suitable concentration and anticoncentration conditions, for a wide range of cost functions.

The contents of this chapter are adapted from the following publication:

- Manole, T., and Niles-Weed, J. (2024). Sharp Convergence Rates for Empirical Optimal Transport with Smooth Costs. *The Annals of Applied Probability*, 34: 1108-1135.

Let us also mention that the results of this chapter have already been extended in several directions. On the one hand, Hundrieser, Staudt, and Munk (2022) showed that the minimax rate in equation (1.5) can be considerably improved if at least one of the measures P and Q has low intrinsic dimension, a phenomenon which they refer to as the *lower complexity adaptation (LCA)* principle. Versions of the LCA principle have also been found to hold for entropic optimal transport (Stromme, 2023; Groppe and Hundrieser, 2023; Pooladian, Divol, and Niles-Weed, 2023). On the other hand, Staudt and Hundrieser (2023) showed that the minimax rate (1.5) continues to hold when the measures P and Q satisfy tail conditions which are considerably weaker than ours. Finally, let us also emphasize that Deb and Mukherjee (2024) have recently identified how the rate (1.5) changes when the underlying samples from P and Q are dependent.

Chapter 4: Sequential Estimation of Optimal Transport Costs.

The previous chapter provides an essentially sharp characterization of the risk of the sample estimator of the optimal transport cost $\mathcal{T}_c(P,Q)$, showing that it typically degrades exponentially with the dimension. Implicit in our proofs, however, is the fact that this rate is driven entirely by the *bias* of the estimator: indeed, it is a simple observation that its variance scales at the parametric rate under mild assumptions. Furthermore, as we already discussed above, it is known that the empirical estimator $\mathcal{T}_c(P_n, Q_n)$ enjoys a $\sqrt{n}$-central limit theorem centered at its expectation, and satisfies the following sub-Gaussian concentration bound (at least when c is bounded): for a constant $C > 0$ and all $\delta \in (0, 1)$,

$$\mathbb{P}\left(\mathcal{T}_c(P_n, Q_n) - \mathbb{E}\mathcal{T}_c(P_n, Q_n) \geq C\sqrt{\frac{\log(1/\delta)}{n}}\right) \leq \delta.$$

The aim of this chapter is to prove that a *sequential* analogue of the above concentration inequality also holds. Specifically, we will prove that,

$$\mathbb{P}\left(\forall n \geq 1 : \mathcal{T}_c(P_n, Q_n) - \mathbb{E}\mathcal{T}_c(P_{\bar{n}}, Q_{\bar{n}}) \geq C\sqrt{\frac{\log(1/\delta) + \log\log n}{n}}\right) \leq \delta,$$

where $\bar{n} = \lceil n/2 \rceil$. Bounds of the above type are sometimes known as *finite law of the iterated logarithm* bounds (Jamieson et al., 2014), and have received a surge of interest in recent years, due to their applications to sequential analysis (see for instance Howard et al. (2021) and references therein). Though our main interest is in finite-sample bounds of the above type, we also show that they lead to an asymptotic one-sided law of the iterated logarithm for optimal transport costs, which may be of independent interest: whenever P is a compactly-supported measure such that $\mathcal{T}_c(P_n, P) = O(n^{-1/2})$ a.s., we show that there is a variance proxy $\sigma^2 > 0$ for which

$$\limsup_{n\to\infty} \frac{n\mathcal{T}_c(P_n, P)}{\sqrt{2\sigma^2 n \log\log n}} \leq 1, \quad \text{almost surely.}$$

Our results are proven using martingale methods, upon noting that the process $\big(\mathcal{T}_c(P_n, Q_n)\big)_{n\geq 1}$ is a reverse submartingale. We also consider the setting where the samples from P and Q are of different sizes, say n and m, in which case we show that $\big(\mathcal{T}_c(P_n, Q_m)\big)_{n,m\geq 1}$ is a partially-ordered reverse submartingale, and distinct techniques need to be adopted.

The contents of this chapter are adapted from the following publication:

- Manole, T., and Ramdas, A. (2023). Martingale Methods for Sequential Estimation of Convex Functionals and Divergences. *IEEE Transactions on Information Theory*, 69:4641–4658.

Chapter 5: Plugin Estimation of Smooth Optimal Transport Maps

In this and the following chapter, we momentarily pause our study of inference for optimal transport costs in order to study the distinct question of estimation and inference for optimal transport maps. It will turn out that our study also leads to several useful tools for the question of inference for the 2-Wasserstein distance, a topic which take up in Chapter 7.

The aim of this chapter is to analyze a number of natural estimators for optimal transport maps, and show that they are minimax optimal. The theoretical analysis of optimal transport map estimators was initiated by Hütter and Rigollet (2021), who derived the minimax rate of estimating optimal transport maps in L^2, under nonparametric smoothness assumptions. Their results show that this problem shares some of the salient features of other nonparametric function estimation problems: optimal transport maps with high smoothness can be estimated at nearly the parametric rate, while those with low smoothness suffer from a curse of dimensionality. Hütter and Rigollet derive an estimator which achieves the minimax rate, but which is computationally intractable. In contrast, in this chapter, we analyze computationally tractable estimators based on the plugin approach: our estimators are simply optimal couplings between measures derived from our observations, appropriately extended so that they define functions on $\mathbb{R}^d$. When the underlying map is assumed to be Lipschitz, we show that computing the optimal coupling between the empirical measures, and extending it using linear smoothers, already gives a minimax optimal estimator. When the underlying map enjoys higher regularity, we show that the optimal coupling between appropriate nonparametric density estimates yields faster rates. Our work also provides new bounds on the risk of corresponding plugin estimators for the quadratic Wasserstein distance, and we show how this problem relates to that of estimating optimal transport maps using stability arguments for smooth and strongly convex Brenier potentials.

The contents of this chapter are adapted from the following preprint:

- Manole, T., Balakrishnan, S., Niles-Weed, J., and Wasserman, L. (2021+). Plugin Estimation of Smooth Optimal Transport Maps. *The Annals of Statistics. (To appear.)*

Chapter 6: Central Limit Theorems for Smooth Optimal Transport Maps

In this chapter, we focus our attention on one of the optimal transport map estimators derived in the previous chapter, and show that it enjoys a pointwise central limit theorem. For simplicity, we limit our setting to that of probability measures supported over the d-dimensional torus $\mathbb{T}^d$, with $d \geq 3$. We analyze a one-sample plugin estimator $\widehat{T}_n$ of the optimal transport map, defined as the unique optimal transport map pushing P forward onto a kernel density estimator of Q with bandwidth $h_n > 0$. Under suitable conditions on the densities and bandwidth, including an undersmoothing condition, we prove that for every point $x \in \mathbb{T}^d$, there is a positive definite matrix $\Sigma(x) \in \mathbb{R}^{d\times d}$ such that

$$\sqrt{nh_n^{d-2}}\big(\widehat{T}_n(x) - T_0(x)\big) \xrightarrow{d} N(0, \Sigma(x)), \quad \text{as } n \to \infty. \tag{1.6}$$

The limiting variance $\Sigma(x)$ can be consistently estimated, thus the above limit law immediately leads to a pointwise confidence band for the true optimal transport map T_0. To the best of our knowledge, this result provides the first procedure for performing inference for the optimal transport map between two absolutely continuous, multivariate distributions (albeit under the strong assumption that the measures are supported on the torus).

The contents of this chapter are adapted from the following preprint:

- Manole, T., Balakrishnan, S., Niles-Weed, J., and Wasserman, L. (2023+). Central Limit Theorems for Smooth Optimal Transport Maps. *arXiv preprint arXiv:2312.12407.*

Chapter 7: Efficient Inference for the Quadratic Wasserstein Distance

In this chapter, we turn back to the study of optimal transport costs, focusing on the quadratic Wasserstein distance. We build upon the ideas developed in the preceding two chapters to develop central limit theorems for estimators of the Wasserstein distance, centered at their population counterpart. These limit laws lead perhaps to the first procedures in the literature for constructing nonparametric confidence intervals for the Wasserstein distance in arbitrary dimension (particularly in the regime $d \geq 4$), assuming the underlying measures are sufficiently regular.

More specifically, we derive limit laws for three classes of estimators. The first are plugin estimators based on suitably-chosen density estimators, of the type introduced in Chapter 5. An appealing aspect of our results is the fact that undersmoothing is not needed: the bandwidth of the density estimators may be taken to be the minimax optimal choice. While this is a useful practical benefit, it has the downside of forcing us to place strong structural assumptions on the underlying domain of the measures, and on the choice of density estimators, in order to guarantee that the resulting plugin estimators do not have leading-order bias. Motivated by this drawback, we introduce a second class of estimators obtained by first-order debiasing. We show that these “one-step” estimators retain the same asymptotics as our plugin estimators, but now the density estimators are permitted to be arbitrary, so long as they satisfy black-box rate conditions. Furthermore, this result removes any assumptions on the underlying support of the measures, an important benefit which we believe could see this estimator gain adoption in practical applications. We close this chapter with a discussion of higher-order debiasing, and show that there exists a second-order debiased estimator over the torus which achieves the same asymptotics as the preceding estimators, but under a significantly weaker smoothness assumption. Nevertheless, as is typical in semiparametric statistics, this second-order estimator loses some of the benefits of its first-order counterpart, since it does not allow for black-box density estimators, and requires us to again place strong structural assumptions on the support of the underlying measures.

In addition to these limit laws, we develop the semiparametric efficiency theory for the Wasserstein distance functional. We state the efficient influence function of the Wasserstein distance, derive asymptotic minimax lower bounds, and show that our various estimators of the Wasserstein distance are asymptotically efficient.

The contents of this chapter are based in part on unpublished work, and in part on results from the paper cited in the above description of Chapter 5.

Chapter 8: An Application to the Search for Pairs of Higgs Bosons

We close this thesis by considering a statistical application of optimal transport to a problem arising in the search for new physical phenomena at the Large Hadron Collider (LHC) of the European Center for Nuclear Research (CERN).

In 2012, the LHC marked its most significant milestone when two independent collaborations announced the discovery of a Higgs boson—a long sought-after particle which was theorized by high energy physicists 50 years prior. Having discovered this Higgs-like particle, current work is concerned with detailed studies of its properties, in order to confirm or refute those predicted by the Standard Model of high energy physics. One such property is the so-called Higgs boson self-coupling, whereby a single excitation of the Higgs field can split into two Higgs bosons without intermediate interactions with other particles. Observing this phenomenon would provide compelling new information regarding the mechanism of particle mass generation, and is therefore considered a major objective for the LHC. Unfortunately, the search for di-Higgs production suffers from a very low signal-to- noise ratio: Collider events which produce pairs of Higgs bosons, known as signal, are indistinguishable from collider events producing other physical processes, known as background. The data arising in this problem is thus a mixture of unlabeled background and signal events, and the goal is to test the null hypothesis that the proportion of signal events is zero. In the high energy physics community, the standard approach to testing these hypotheses treats the background distribution as a nuisance parameter, which must first be estimated using held-out data. This is known as the problem of *data-driven background estimation*.

This chapter proposes a new approach to the data-driven background estimation problem by casting it as a domain adaptation problem. We then show how optimal transport maps can be used to correct the underlying distribution shift. Our method relies on modeling assumptions which are complementary to those used by existing approaches in experimental high energy physics. It can thus serve as a powerful cross-check for existing methods, an important benefit which we hope will increase the analyst's trust in the obtained background estimates of future studies at the LHC. Our approach involves the estimation of optimal transport maps, for which we use some of the estimators introduced in Chapter 5.

The contents of this chapter are adapted from the following preprint:

- Manole, T., Bryant, P., Alison, J., Kuusela, M., and Wasserman, L. (2022+). Background Modeling for Double Higgs Boson Production: Density Ratios and Optimal Transport. *Under Minor Revision, The Annals of Applied Statistics. arXiv preprint arXiv:2208.02807.*

Additional Contributions

During my PhD, I also pursued a separate line of work that does not appear in this thesis, in which I used optimal transport as a theoretical tool for the analysis of estimation and model selection procedures in finite mixture models. This research has appeared in the following preprints and publications:

- Manole, T., Ho, N. (2020+). "Uniform Convergence Rates for Maximum Likelihood Estimation under Two-Component Finite Mixture Models". *arXiv preprint arXiv:2006.00704.*
- Manole, T., Khalili, A. (2021). "Estimating the Number of Components in Finite Mixture Models via the Group-Sort-Fuse Procedure". *The Annals of Statistics 49, 3043–3069.*
- Manole, T., Ho, N. (2022). "Refined Convergence Rates for Maximum Likelihood Estimation in Finite Mixtures Models". *Proceedings of the 39th International Conference on Machine Learning, PMLR 162:14979-15006.*

1.3 Background on Optimal Transport

Before turning to the opening chapter of this thesis, we provide a brief introduction to the notions from optimal transport which will be used throughout what follows. All notation introduced in this section will be used systematically in subsequent chapters, and all additional notational conventions are summarized in Section 1.4 below.

1.3.1 Generalities

The Monge and Kantorovich Problems Let $d \geq 1$ be an integer, and let $\mathcal{X}$ and $\mathcal{Y}$ denote subsets[1] of $\mathbb{R}^d$. Let $\mathbb{B}(\mathcal{X})$ denote the Borel σ-algebra over $\mathcal{X}$, and let $\mathcal{P}(\mathcal{X})$ denote the set of Borel probability measures over $\mathcal{X}$. Given two probability measures $P \in \mathcal{P}(\mathcal{X})$ and $Q \in \mathcal{P}(\mathcal{Y})$, a map $T : \mathcal{X} \to \mathcal{Y}$ is said to push P forward onto Q if it satisfies

$$T_\# P(B) := P(T^{-1}(B)) = Q(B), \quad \text{for all } B \in \mathbb{B}(\mathcal{Y}).$$

The above display is equivalent to the requirement that $X \sim P$ imply $T(X) \sim Q$. Whenever this property holds, we say that T is a *transport map* between P and Q, and the set of all such maps is denoted $\mathcal{T}(P, Q)$.

Let $c : \mathcal{X} \times \mathcal{Y} \to \mathbb{R}$ be a nonnegative Borel-measurable function, known as a *cost function*. The *Monge optimal transport problem* is then defined as the optimization problem

$$\inf_{T \in \mathcal{T}(P,Q)} \int_{\mathcal{X}} c(x, T(x)) dP(x). \tag{1.7}$$

Whenever a minimizer exists, we shall denote it by $T_0 \in \mathcal{T}(P, Q)$. The solvability of the Monge problem is, however, a delicate question. The set $\mathcal{T}(P, Q)$ may generally be empty, which is for instance the case whenever $P = \delta_x$ is a Dirac mass at a point $x \in \mathcal{X}$, and Q is supported on at least two points in $\mathcal{Y}$. Even when $\mathcal{T}(P, Q)$ is nonempty, solutions the minimization problem (1.7) are generally difficult to analyze due to the lack of convexity of both the objective function and of the set $\mathcal{T}(P, Q)$—we refrain from saying more for now, and refer the reader to Gangbo and McCann (1996) and Villani (2008, Chapter 9) for sufficient conditions on P, Q, c under which unique solutions of the Monge problem can be derived.

These difficulties motivated Kantorovich (1942, 1948) to develop a convex relaxation of the Monge problem. This relaxation will involve an optimization over *couplings* rather than transport maps. A coupling between P and Q is a joint distribution $\pi \in \mathcal{P}(\mathcal{X} \times \mathcal{Y})$ with marginal distributions P and Q, and the set of such couplings is denoted

$$\Pi(P, Q) = \left\{ \pi \in \mathcal{P}(\mathcal{X} \times \mathcal{Y}) : \pi(\cdot \times \mathcal{Y}) = P, \pi(\mathcal{X} \times \cdot) = Q \right\}.$$

The *Kantorovich problem* is then defined as

$$\inf_{\pi \in \Pi(P,Q)} \int_{\Omega} c(x, y) d\pi(x, y). \tag{1.8}$$

[1] Whenever $\mathcal{X} = \mathcal{Y}$ in subsequent chapters, we typically write $\Omega := \mathcal{X} = \mathcal{Y}$. The only exception is Chapter 4, where we prefer to retain the notation Ω to denote the sample space of the underlying probability space.

Unlike the Monge problem, the Kantorovich problem is always feasible since $P \otimes Q \in \Pi(P, Q)$. Furthermore, a minimizer π_0 in equation (1.8) exists as soon as the cost function c is lower semi-continuous (Villani (2008), Theorem 4.1), and is called an optimal coupling. In the special case where π_0 is supported in the graph of a map $T_0 : \mathcal{X} \to \mathcal{Y}$, namely on a set of the form $\{(x, T_0(x)) : x \in \mathcal{X}\}$, it must be the case that $T_0 \in \mathcal{T}(P, Q)$ due to the marginal constraints in the definition of $\Pi(P, Q)$, and it must then follow that T_0 is an optimal transport map from P to Q. Therefore, the Kantorovich problem is indeed a relaxation of the Monge problem. Crucially, this relaxation is a convex (albeit infinite-dimensional) problem.

The Kantorovich Duality In addition to being convex, the Kantorovich problem is a linear program. It admits a dual maximization problem, known as the *Kantorovich dual problem*, defined by

$$\sup_{(\phi,\psi)\in\Phi_c(P,Q)} \int \phi dP + \int \psi dQ, \tag{1.9}$$

where

$$\begin{aligned}\Phi_c(P, Q) = \Big\{&(\phi, \psi) \in L^1(P) \times L^1(Q) : \\ &\qquad \phi(x) + \psi(y) \leq c(x, y) \text{ for } P\text{-a.e. } x \in \mathcal{X} \text{ and } Q\text{-a.e. } y \in \mathcal{Y}\Big\}.\end{aligned} \tag{1.10}$$

As soon as c is lower semicontinuous, it can be shown that strong duality holds, so that the optimal objective values in equations (1.8) and (1.9) are equal. Furthermore, under a mild integrability condition on c, the maximizer of the dual Kantorivich problem admits a solution $(\phi_0, \psi_0) \in \Phi_c(P, Q)$ (Villani (2008), Theorem 5.10), known as a pair of *Kantorovich potentials*.

It is useful to note that the description of the set $\Phi_c(P, Q)$ may be significantly simplified. Indeed, once a solution $(\phi_0, \psi_0) \in \Phi_c(P, Q)$ exists, notice that the pair (ϕ_0, ϕ_0^c), where

$$\phi_0^c(y) = \inf_{x\in\mathcal{X}} \{c(x, y) - \phi_0(x)\}, \tag{1.11}$$

is itself a pair of Kantorovich potentials, since replacing ψ_0 by ϕ_0^c can only increase the objective value (1.9), while retaining the constraint $(\phi_0, \phi_0^c) \in \Phi_c(P, Q)$. ϕ_0^c is called the *c-conjugate* of ϕ_0, and any function on $\mathcal{Y}$ which can be written as f^c for some $f : \mathcal{X} \to \overline{\mathbb{R}} := \mathbb{R} \cup \{-\infty\}$ is said to be *c-concave*, provided that it is not identically equal to $-\infty$. It can be deduced that the Kantorovich dual problem is equivalent to

$$\sup_{\phi\in L^1(P)} \int \phi dP + \int \phi^c dQ, \tag{1.12}$$

where the supremum can further be restricted to the set of c-concave functions in $L^1(P)$.

Optimal Transport Costs and Wasserstein Distances The main quantities of interest in the Monge and Kantorovich problems are the optimal transport map T_0, or the optimal coupling π_0, as they define an optimal transference plan between the probability measures P and Q. Of equal interest is the cost of this optimal transference plan, as it provides a notion of divergence between P and Q. We define the *optimal transport cost* as the optimal objective value in the Kantorovich problem, that is,

$$\mathcal{T}_c(P,Q) = \inf_{\pi \in \Pi(P,Q)} \int_{\mathcal{X}\times\mathcal{Y}} c(x,y) d\pi(x,y). \tag{1.13}$$

Optimal transport costs can be used to define a family of metrics between probability measures known as the Wasserstein distances. Assume $\Omega := \mathcal{X} = \mathcal{Y}$, let $r \geq 1$, and define

$$\mathcal{P}_r(\mathcal{X}) = \left\{ P \in \mathcal{P}_r(\Omega) : \int_\Omega \|x\|^r dP(x) < \infty \right\}.$$

Then, the r-Wasserstein distance is defined as

$$W_r : \mathcal{P}_r^2(\Omega) \to \mathbb{R}_+, \quad W_r = \mathcal{T}_{\|\cdot\|^r}^{1/r}.$$

Wasserstein distances can naturally be interpreted as measuring the cost of deforming one measure into another. Furthermore, they are sensitive to the topology of the underlying space Ω due to the metric d embedded into their definition. These are important motivations for their use in statistical applications. We also note that Wasserstein distances place no assumptions on the absolute continuity of P with respect to Q. Therefore, unlike many classical metrics between probability distributions, plugin estimators for Wasserstein distances are well-defined without any smoothing.

1.3.2 The Quadratic Optimal Transport Problem over $\mathbb{R}^d$

We shall pay particular attention to the optimal transport problem with respect to the quadratic cost $c(x,y) = \|x-y\|^2$. This case is arguably most important for statistical applications, and turns out to be simplest to analyze due to a celebrated result of Brenier which we have already informally seen in Theorem 1: a coupling is optimal for the quadratic cost if and only if it is supported on the subdifferential of a convex function. Let us explore this result in further detail.

Brenier's Theorem Let $P \in \mathcal{P}_2(\mathcal{X})$ and $Q \in \mathcal{P}_2(\mathcal{Y})$. Since $\|\cdot\|^2$ is smooth, strong duality holds in the Kantorovich dual problem with respect to the quadratic cost, and there exists a pair of Kantorovich potentials (ϕ_0, ϕ_0^c) solving the maximization problem (1.12). If we define $\varphi_0 = \|\cdot\|^2 - 2\phi_0$, then ϕ_0^c equivalently takes the form $\phi_0^c = \|\cdot\|^2 - 2\varphi_0^*$, where for any function $f : \mathbb{R}^d \to \mathbb{R} \cup \{\infty\}$, we denote by

$$f^*(y) = \sup_{x \in \mathbb{R}^d} \big\{ \langle x, y\rangle - f(x) \big\}, \quad y \in \mathbb{R}^d,$$

the Legendre-Fenchel conjugate of f. Under this reparametrization, the Kantorovich dual problem is equivalent to the so-called semi-dual problem

$$\inf_{\varphi \in L^1(P)} \int \varphi dP + \int \varphi^* dQ, \tag{1.14}$$

in the sense that φ_0 is a solution to the semi-dual problem if and only if $(\|\cdot\|^2 - 2\varphi_0, \|\cdot\|^2 - 2\varphi_0^*)$ is a solution to the Kantorovich dual problem (1.9). Solutions to the semi-dual problem are closely related to the Monge problem, as described by the following result due to Knott and Smith (1984); Brenier (1991). We denote by $\mathcal{P}_{\mathrm{ac}}(\mathcal{X})$ the set of measures in $\mathcal{P}(\mathcal{X})$ which are absolutely continuous with respect to the Lebesgue measure on $\mathcal{X}$.

Theorem 2 (Brenier's Theorem)**.** Let $P \in \mathcal{P}_{\mathrm{ac}}(\Omega)$ and $Q \in \mathcal{P}(\Omega)$.

(i) There exists an optimal transport map T_0 between P and Q which takes the form $T_0 = \nabla\varphi_0$ for a convex function $\varphi_0 : \mathbb{R}^d \to \mathbb{R}$ which solves the semi-dual problem (1.14). Furthermore, T_0 is uniquely determined P-almost everywhere.

(ii) If we further have $Q \in \mathcal{P}_{\mathrm{ac}}(\Omega)$, then $\nabla\varphi_0^*$ is the (Q-almost everywhere uniquely determined) gradient of a convex function such that $\nabla\varphi^*_{0\#}Q = P$, and solves the Monge problem for transporting Q onto P. Furthermore, for Lebesgue-almost every $x, y \in \Omega$

$$\nabla\varphi_0^* \circ \nabla\varphi_0(x) = x, \quad \nabla\varphi_0 \circ \nabla\varphi_0^*(y) = y.$$

Brenier's Theorem implies that a unique optimal transport map exists between any absolutely continuous distribution P and any distribution Q, where uniqueness is always understood in the Lebesgue-almost everywhere sense. It further characterizes this map as the gradient of an optimal semi-dual potential φ_0, which we also refer to as a *Brenier potential.* Unlike optimal transport maps, we emphasize that Brenier potentials are not a.e.-uniquely determined by P and Q; for instance, one may always add a constant to φ_0 without changing its gradient.

Regularity of Optimal Transport Maps Let $\Omega := \mathcal{X} = \mathcal{Y}$, and let the distributions $P, Q \in \mathcal{P}_{\mathrm{ac}}(\Omega)$ admit respective Lebesgue densities p, q over some compact set $\Omega \subseteq \mathbb{R}^d$. Fix a Brenier potential φ_0 whose gradient T_0 pushes forward P onto Q. The convexity of φ_0 implies that it will be almost-everywhere twice differentiable. The basic question we will try to answer in this section is the following: what are sufficient conditions on P, Q, Ω which ensure that T_0 has entries lying in the Hölder space $\mathcal{C}^\alpha(\Omega)$, for some $\alpha > 0$?

Regularity properties of Brenier potentials, and therefore of optimal transport maps, have historically been studied via the regularity theory of partial differential equations of the Monge-Ampère type; we refer to De Philippis and Figalli (2014); Figalli (2017) for surveys. To understand this connection, notice that if φ_0 were in fact *everywhere* twice continuously differentiable, then the constraint $\nabla\varphi_{0\#}P = Q$ would imply—by the traditional change-of-variable formula—that φ_0 solves the equation

$$\det\left(\nabla^2\varphi_0(x)\right) = \frac{p(x)}{q(\nabla\varphi_0(x))}, \quad x \in \Omega. \tag{1.15}$$

Equation (1.15) is a second-order, fully nonlinear elliptic PDE which falls into the class of *Monge-Ampère* equations. By analogy with the regularity theory of uniformly elliptic PDEs of second-order (Gilbarg and Trudinger, 2001), it is natural to guess from equation (1.15) that φ_0 generally admits two degrees of smoothness more than the densities p and q, provided that q is bounded away from zero. This intuition indeed turns out to hold true under suitable regularity conditions on Ω, as was established in a series of publications by Caffarelli (1991), Caffarelli (1992, 1992), Caffarelli (1996). The following is a summary of these results, as stated by Villani (2008, Chapter 12).

Theorem 3 (Caffarelli's Regularity Theory)**.** Assume Ω is a compact, convex set. Assume further that there exists $\gamma > 0$ such that $\gamma^{-1} \leq p, q \leq \gamma$ over Ω. Then, the Brenier potential φ_0 is unique up to an additive constant, and satisfies the following.

(i) (Interior Regularity) Suppose there exists $\alpha > 1$, $\alpha \notin \mathbb{N}$, such that $p, q \in \mathcal{C}^{\alpha-1}(\Omega^\circ)$. Then $\varphi_0 \in \mathcal{C}^{\alpha+1}(\Omega^\circ)$. Moreover, for any open subdomain Ω' such that $\overline{\Omega'} \subseteq \Omega^\circ$, there exists a constant $C > 0$ depending on $\gamma, \alpha, \Omega, \Omega'$, $\|\varphi_0\|_{L^\infty(\Omega)}$, $\|p\|_{\mathcal{C}^{\alpha-1}(\Omega^\circ)}$, $\|q\|_{\mathcal{C}^{\alpha-1}(\Omega^\circ)}$ such that

$$\|\varphi_0\|_{\mathcal{C}^{\alpha+1}(\Omega')} \leq C.$$

(ii) (Global Regularity) Assume Ω admits a $\mathcal{C}^2$ boundary and is uniformly convex. Assume further that there exists $\alpha > 1$, $\alpha \notin \mathbb{N}$, such that $p, q \in \mathcal{C}^{\alpha-1}(\Omega)$. Then, $\varphi_0 \in \mathcal{C}^{\alpha+1}(\Omega)$.

Theorem 3(ii) implies that, under suitable conditions, the optimal transport map T_0 inherits one degree of smoothness more than the densities p and q over Ω. Unlike the interior regularity result of Theorem 3(i), however, Theorem 3(ii) does not imply a uniform bound on $\|\varphi_0\|_{\mathcal{C}^{\alpha+1}(\Omega)}$, and therefore does not preclude the possibility that the latter quantity diverges when p, q vary in a $\mathcal{C}^{\alpha-1}(\Omega)$ ball. Closely related global regularity results have also been established by Urbas (1997) under slightly stronger conditions, but we do not know if either of these results can be made uniform up to the boundary in an analogous way to the interior result of Theorem 3(i). Whenever global uniform regularity results are needed in our development, we sidestep this issue by working with the optimal transport problem over the torus, for which boundary considerations do not arise.

1.3.3 The Quadratic Optimal Transport Problem over the Flat Torus

Denote by $\mathbb{T}^d = \mathbb{R}^d/\mathbb{Z}^d$ the flat d-dimensional torus. Specifically, $\mathbb{T}^d$ is the set of equivalence classes $[x] = \{x + k : k \in \mathbb{Z}^d\}$, for all $x \in [0, 1)^d$. Abusing notation, we typically write x instead of $[x]$. $\mathbb{T}^d$ is endowed with the standard metric[2]

$$\|x - y\|_{\mathbb{T}^d} = \min\{\|x - y + k\| : k \in \mathbb{Z}^d\}, \quad x, y \in \mathbb{T}^d.$$

We identify $\mathcal{P}(\mathbb{T}^d)$ with the set of Borel measures P on $\mathbb{R}^d$ such that $P([0, 1)^d) = 1$ and which are $\mathbb{Z}^d$-periodic, in the sense that $P(B) = P(k + B)$ for all $k \in \mathbb{Z}^d$ and all Borel sets $B \subseteq \mathbb{R}^d$. Furthermore, $\mathcal{P}_{\text{ac}}(\mathbb{T}^d)$ denotes the subset of measures in $\mathcal{P}(\mathbb{T}^d)$ which are absolutely continuous with respect to the Lebesgue measure on $\mathbb{R}^d$. A function $f : \mathbb{T}^d \to \mathbb{R}$ is understood to be a function on $\mathbb{R}^d$ which is $\mathbb{Z}^d$-periodic, and we write $T : \mathbb{T}^d \to \mathbb{T}^d$ when T is a map from $\mathbb{R}^d$ to $\mathbb{R}^d$ such that $[T(x)] = [T(y)]$ whenever $[x] = [y]$.

[2] Of course, $\| \cdot - \cdot \|_{\mathbb{T}^d}$ does not define a norm; this notation is meant to be taken at face value. We write $\|x\|_{\mathbb{T}^d} := \|x - 0\|_{\mathbb{T}^d}$ for any $x \in \mathbb{T}^d$.

The optimal transport problem over $\mathbb{T}^d$ with the quadratic cost $d^2_{\mathbb{T}^d}$ largely mirrors that of the squared Euclidean cost over $\mathbb{R}^d$. Define for all $P, Q \in \mathcal{P}_{\mathrm{ac}}(\mathbb{T}^d)$ the Monge problem

$$\operatorname*{argmin}_{T \in \mathcal{T}(P,Q)} \int_{\mathbb{T}^d} \|x - T(x)\|^2_{\mathbb{T}^d} dP(x), \tag{1.16}$$

where the integral is understood as being taken over $[0,1)^d$. The Kantorovich problem and its dual give rise to the squared Wasserstein distance over $\mathcal{P}(\mathbb{T}^d)$,

$$\mathcal{W}_2^2(P,Q) = \inf_{\pi \in \Pi(P,Q)} \int \|x - y\|^2_{\mathbb{T}^d} d\pi(x,y) = \sup_{(\phi,\psi) \in \mathcal{K}_T} \int \phi dP + \int \psi dQ, \tag{1.17}$$

where $\mathcal{K}_T$ denotes the set of pairs of potentials $(\varphi, \psi) \in L^1(P) \times L^1(Q)$ satisfying the dual constraint $\varphi(x) + \psi(y) \le \|x - y\|^2_{\mathbb{T}^d}$ for all $x, y \in \mathbb{T}^d$. We abuse notation by writing W_2 to denote both the 2-Wasserstein distance over $\mathbb{R}^d$ and $\mathbb{T}^d$. Whenever we speak of the optimal transport problem or Wasserstein distance between two measures $P, Q \in \mathcal{P}(\Omega)$, the underlying cost function is tacitly understood to be $\|\cdot - \cdot\|^2$ when $\Omega \subseteq \mathbb{R}^d$, and $\|\cdot - \cdot\|^2_{\mathbb{T}^d}$ when $\Omega = \mathbb{T}^d$.

The following result due to Cordero-Erausquin (1999) is an analogue of Brenier's Theorem, together with additional properties about the optimal transport problem over $\mathbb{T}^d$.

Proposition 1. Let $P \in \mathcal{P}_{\mathrm{ac}}(\mathbb{T}^d)$ and $Q \in \mathcal{P}(\mathbb{T}^d)$. Then, there exists a ($P$-a.e. uniquely determined) optimal transport map $T_0 = \nabla\varphi_0$ from P to Q which solves the Monge problem (1.16), where $\varphi_0 : \mathbb{R}^d \to \mathbb{R}$ is a convex function satisfying the following properties.

(i) $\|\cdot\|^2 /2 - \varphi_0$ is $\mathbb{Z}^d$-periodic.

(ii) $T_0(x + k) = T_0(x) + k$ for almost every $x \in \mathbb{R}^d$ and $k \in \mathbb{Z}^d$.

(iii) For P-almost all $x \in \mathbb{R}^d$, $\|T_0(x) - x\| = d_{\mathbb{T}^d}(x, T_0(x))$.

Assume further that $Q \in \mathcal{P}_{\mathrm{ac}}(\mathbb{T}^d)$, and denote the respective densities of P, Q by p, q. Then,

(v) $\nabla\varphi_0^*$ is the (Q-a.e. uniquely determined) optimal transport map from Q to P.

(vi) $(\|\cdot\|^2 - 2\varphi_0, \|\cdot\|^2 - 2\varphi_0^*)$ is a pair of optimal Kantorovich potentials in equation (1.17).

(vii) If $\varphi_0 \in \mathcal{C}^2([0,1]^d)$, then it solves the Monge-Ampère equation

$$\det(\nabla^2\varphi_0(x))q(\nabla\varphi_0(x)) = p(x), \quad x \in \mathbb{R}^d.$$

In particular, if $\gamma^{-1} \le p, q \le \gamma$ for some $\gamma > 0$, then φ_0 is λ-strongly convex, for some constant $\lambda > 0$ depending only on γ and $\|\varphi_0\|_{\mathcal{C}^2([0,1]^d)}$.

With Proposition 1 in place, regularity properties of Brenier potentials φ_0 may be deduced from smoothness conditions on p, q. The following result was stated by Cordero-Erausquin (1999) without explicit mention of the uniformity of the Hölder norms appearing therein, but can readily be made uniform using Caffarelli's interior regularity theory (Theorem 3(i); Figalli (2017), Chapter 4). We also note that this result was stated by Ambrosio et al. (2012) in the special case $d = 2$.

Theorem 4. Let $P, Q \in \mathcal{P}(\mathbb{T}^d)$ be absolutely continuous with respect to the Lebesgue measure, with respective densities p, q satisfying $\gamma^{-1} \leq p, q \leq \gamma$ for some $\gamma > 0$. Assume further that $p, q \in \mathcal{C}^{\alpha-1}(\mathbb{T}^d)$ for some $\alpha > 1$. Then, there exists a constant $C > 0$ depending only on $\alpha, \gamma, \|p\|_{\mathcal{C}^{\alpha-1}(\mathbb{T}^d)}$ and $\|q\|_{\mathcal{C}^{\alpha-1}(\mathbb{T}^d)}$ such that, $\|\varphi_0\|_{\mathcal{C}^{\alpha+1}([0,1]^d)} \leq C$.

1.4 Notation

The following notational conventions will be used throughout this thesis.

Elementary Definitions. The Euclidean norm on $\mathbb{R}^d$ is denoted $\|\cdot\|$, and the ℓ_p norm of a sequence $(a_n)_{n\geq 1} \subseteq \mathbb{R}$ is written $\|(a_n)_{n\geq 1}\|_{\ell_p} = (\sum_{n\geq 1} |a_n|^p)^{1/p}$ for all $1 \leq p \leq \infty$. For any $B \in \mathbb{N}$, the permutation group on $[B] = \{1, \dots, B\}$ is denoted S_B. For any $a, b \in \mathbb{R}$, let $a \vee b = \max\{a, b\}$ and $a \wedge b = \min\{a, b\}$, $a_+ = a \vee 0$. If $a \geq 0$, $\lfloor a \rfloor$ and $\lceil a \rceil$ denote the respective floor and ceiling of a. We write $\overline{\mathbb{R}} = \mathbb{R} \cup \{\infty\}, \mathbb{R}_+ = [0, \infty), \mathbb{N} = \{1, 2, \dots\}$ and $\mathbb{N}_0 = \{0, 1, \dots\}$. Given a set $\Omega \subseteq \mathbb{R}^d$, its interior, closure, and boundary with respect to the Euclidean norm are respectively denoted Ω°,$\overline{\Omega}$, and $\partial\Omega$. The logarithm of base $b > 1$ is denoted $\log_b(x) = \log x / \log b$, where $\log$ is the natural logarithm. For all $x \in \mathbb{R}^d$ and $\epsilon > 0$, $B(x, \epsilon) \equiv B_{x,\epsilon} = \{y \in \mathbb{R}^d : \|x - y\| \leq \epsilon\}$.

Given a twice differentiable map $f : \mathbb{R}^d \to \mathbb{R}$, the gradient of f is denoted ∇f, its Hessian is denoted $\nabla^2 f$, and its Laplacian is denoted $\Delta f = \sum_{i=1}^d \partial^2 f / \partial x_i^2$. The divergence of a differentiable vector field $F = (F_1, \dots, F_d) : \mathbb{R}^d \to \mathbb{R}^d$ is denoted $\operatorname{div}(F) = \sum_{i=1}^d \partial F_i / \partial x_i$.

Given a measure space $(\Omega, \mathcal{F}, \mathbb{P})$, and two sub-$\sigma$-algebras $\mathcal{G}, \mathcal{H} \subseteq \mathcal{F}$, we use the standard notation for joins and intersections of σ-algebras, respectively given by $\mathcal{G} \bigvee \mathcal{H} := \sigma(\mathcal{G} \cup \mathcal{H})$ and $\mathcal{G} \bigwedge \mathcal{H} := \mathcal{G} \cap \mathcal{H}$.

Convex Analysis. Given a convex function $\varphi : \mathbb{R}^d \to \mathbb{R} \cup \{\infty\}$, we denote by

$$\varphi^*(y) = \sup_{x \in \mathbb{R}^d} \Big\{ \langle x, y \rangle - \varphi(x) \Big\}, \quad y \in \mathbb{R}^d$$

its convex conjugate, and by

$$\partial\varphi(x) = \Big\{ z \in \mathbb{R}^d : \varphi(y) - \varphi(x) \geq \langle z, y - x \rangle, \ y \in \mathbb{R}^d \Big\}$$

its subdifferential at a point $x \in \mathbb{R}^d$.

Spaces of Probability Measures. Given a set $\Omega \subseteq \mathbb{R}^d$, we denote by $\mathcal{P}(\Omega)$ the set of Borel probability measures on Ω, and for any $r > 0$,

$$\mathcal{P}_r(\Omega) = \left\{ P \in \mathcal{P}(\Omega) : \int_\Omega \|x\|^r dP(x) < \infty \right\}.$$

Furthermore, $\mathcal{P}_{\mathrm{ac}}(\Omega)$ denotes the subset of measures in $\mathcal{P}(\Omega)$ which are absolutely continuous with respect to the Lebesgue measure $\mathcal{L}$. We adopt similar conventions when Ω is replaced by the flat torus $\mathbb{T}^d = \mathbb{R}^d/\mathbb{T}^d$. A probability measure $P \in \mathcal{P}(\mathbb{R}^d)$ is said to be σ^2-sub-Gaussian for some $\sigma^2 > 0$ if it holds for $Y \sim P$ that

$$\mathbb{E}\exp(\lambda^\top(Y - \mathbb{E}Y)) \leq \exp(\|\lambda\|^2 \sigma^2/2), \quad \text{for all } \lambda \in \mathbb{R}^d.$$

Divergences between Probability Measures. Let $P, Q \in \mathcal{P}(\mathbb{R}^d)$ be probability measures such that $P \ll Q$. We denote by

$$\|P - Q\|_{\mathrm{TV}} = \sup_{A \in \mathcal{B}(\mathbb{R}^d)} \big|P(A) - Q(A)\big|, \quad h^2(P, Q) = \int \left(\sqrt{dP} - \sqrt{dQ}\right)^2$$

the total variation and Hellinger distances, and by

$$\mathrm{KL}(P, Q) = \int \log\left(\frac{dP}{dQ}\right) dP, \quad \chi^2(P, Q) = \int \left(\frac{dP}{dQ} - 1\right)^2 dQ,$$

the Kullback-Leibler and χ^2-divergences.

Function Spaces. Given a measure space $(\Omega, \mathcal{F}, \nu)$, $L^p(\nu)$ denotes the Lebesgue space of order $1 \leq p \leq \infty$, endowed with the norm $\|f\|_{L^p(\nu)} = (\int_\Omega |f(x)|^p d\nu(x))^{1/p}$, for any measurable function $f : \Omega \to \mathbb{R}$. Throughout this thesis we will adopt the following nonstandard convention: the subscript "0" on a function space will always be used to indicate a restriction to mean zero functions, and will never refer to a boundary condition. For example, we write

$$L_0^p(\nu) = \{f \in L^p(\nu) : \int f d\nu = 0\}.$$

When ν is the Lebesgue measure $\mathcal{L}$ on $\Omega \subseteq \mathbb{R}^d$, we write $L^p(\Omega)$ (or $L_0^p(\Omega)$) instead of $L^p(\mathcal{L})$ (or $L_0^p(\mathcal{L})$). We adopt the same convention when Ω is equal to the flat torus $\mathbb{T}^d = \mathbb{R}^d/\mathbb{Z}^d$, in which case, by abuse of notation, $\mathcal{L}$ denotes the standard Haar measure over $\mathbb{T}^d$, normalized to be a probability measure. We often abbreviate $\int f d\mathcal{L}$ by $\int f$.

Given a map $T = (T_1, \dots, T_d) : \Omega \to \Omega$, where Ω is either a subset of $\mathbb{R}^d$ or the flat torus $\mathbb{T}^d$, we will write by abuse of notation

$$\|T\|_{\mathcal{C}^s(\Omega)} := \sum_{i=1}^d \|T_i\|_{\mathcal{C}^s(\Omega)}, \quad \|T\|_{L^r(\Omega)} := \sum_{i=1}^d \|T_i\|_{L^r(\Omega)},$$

for any $s > 0, 1 \leq r \leq \infty$.

In defining standard function spaces, we follow similar conventions as Triebel (1983); Schmeisser and Triebel (1987). In particular, given a set Ω, which is either a closed subset of $\mathbb{R}^d$ or the d-dimensional flat torus $\Omega = \mathbb{T}^d$, and given real numbers $\alpha > 0$, $s \in \mathbb{R} \setminus \{0\}$, $1 \leq p, q \leq \infty$, the Hölder spaces $\mathcal{C}^\alpha(\Omega)$, Besov spaces $\mathcal{B}^s_{p,q}(\Omega)$, Sobolev spaces $\dot{H}^{s,p}(\Omega)$, and their respective norms $\|\cdot\|_{\mathcal{C}^\alpha(\Omega)}$, $\|\cdot\|_{\mathcal{B}^s_{p,q}(\Omega)}$, $\|\cdot\|_{H^{s,r}(\Omega)}$, are defined in Appendix A. We drop the suffix Ω when the underlying space can be understood from context.

Fourier Transform and Convolution. The Fourier transform of a function $K \in L^1(\mathbb{R}^d)$ is denoted

$$\mathcal{F}[K](\xi) = \int_{\mathbb{R}^d} f(x) e^{-2\pi i x^\top \xi} dx, \quad \text{for any } \xi \in \mathbb{R}^d.$$

Given a map $f \in L^2(\mathbb{T}^d)$, we continue to denote by $\mathcal{F}[f](\xi)$ the Fourier coefficients of f, now restricted to all $\xi \in \mathbb{Z}^d$. We also write $\mathbb{Z}^d_* = \mathbb{Z}^d$.

The convolution of functions $f, g : \mathbb{R}^d \to \mathbb{R}$ is denoted by $(f \star g)(x) = \int f(x-y)g(y)dy$. Furthermore, the convolution of two Borel probability measures P, Q is the measure $(P \star Q)(B) = \int I_B(x+y)dP(x)dQ(y)$ for all $B \in \mathbb{B}(\mathbb{R}^d)$, where $I_B(x) = I(x \in B)$ is the indicator function of B. The convolution of P with f is the function $(P \star f)(x) = \int f(x-y)dP(y)$.

Constants. Given two real numbers $a, b > 0$, we write $a \lesssim b$ if there exists a universal constant $C > 0$—not depending on a and b but possibly depending on quantities which are either clear from context or specified explicitly—such that $a \leq Cb$. We write $a \asymp b$ if $a \lesssim b \lesssim a$. When we wish to emphasize the dependence of C on a particular quantity g, we may write $\lesssim_g$ or $\asymp_g$.

Many constants throughout Chapters 5–7 will depend on quantities such as $\|f\|_{\mathcal{C}^s(\Omega)}$ and $\|f^{-1}\|_{L^\infty(\Omega)}$, for some $s > 0$ and $f \in \mathcal{C}^s(\Omega)$. We therefore introduce the following abbreviation: for any $s > 0$, $k \geq 1$, $f_1, \dots, f_k \in \mathcal{C}^s_+(\mathbb{T}^d)$,

$$\omega_s(f_1, f_2, \dots, f_k) := \sum_{j=1}^k \left(\|f_j\|_{\mathcal{C}^s(\mathbb{T}^d)} + \|f_j^{-1}\|_{L^\infty(\mathbb{T}^d)} \right). \tag{1.18}$$

Part I

Empirical Optimal Transport

Chapter 2

Minimax Confidence Intervals for the Sliced Wasserstein Distance

2.1 Introduction

In this chapter, we take a first step toward the question of performing statistical inference for the Wasserstein distance, by studying a related functional known as the Sliced Wasserstein distance. Introduced by Rabin et al. (2011); Bonneel et al. (2015), this metric is obtained by averaging the Wasserstein distance between random one-dimensional projections of the distributions to be compared. Concretely, let $\mathbb{S}^{d-1}$ be the unit ball of $\mathbb{R}^d$ with respect to the Euclidean norm, and let μ denote the uniform probability measure on $\mathbb{S}^{d-1}$. Then, the r-th order Sliced Wasserstein distance between two probability distributions $P, Q \in \mathcal{P}_r(\mathbb{R}^d)$, for some $r \geq 1$, is given by

$$\mathrm{SW}_r(P, Q) = \left(\int_{\mathbb{S}^{d-1}} W_r^r(P_\theta, Q_\theta) d\mu(\theta) \right)^{\frac{1}{r}}, \tag{2.1}$$

where for any $\theta \in \mathbb{S}^{d-1}$, we let $\pi_\theta : x \in \mathbb{R}^d \mapsto x^\top \theta$, and write $P_\theta = \pi_{\theta\#}P$ and $Q_\theta = \pi_{\theta\#}Q$.

The primary motivation for defining the Sliced Wasserstein distance is its computational tractability. Indeed, in spite of the recent popularity of the Wasserstein distance, its high computational complexity often limits its applicability to large-scale problems. A key exception to this high computational cost is the univariate case, in which the Wasserstein distance admits a closed form as the L^r norm between the quantile functions of the distributions to be compared, which is easy to compute. Since the measures P_θ and Q_θ appearing in the above display are univariate, it follows that the Sliced Wasserstein is itself computable in closed form (as we shall see in Section 2.2). Furthermore, the Sliced Wasserstein distance shares some of the same qualitative behaviour as the multivariate Wasserstein distance, and even induces the same topology on $\mathcal{P}_r(\mathbb{R}^d)$ (Bonnotte, 2013). These considerations make it an attractive and easily computable alternative to the Wasserstein distance in large-scale applications.

Motivated by the fact that the Wasserstein distance and its sliced analogue are sensitive to outliers and heavy tails, we introduce a trimmed version of the Sliced Wasserstein distance, denoted by $\mathrm{SW}_{r,\delta}(P,Q)$ for some trimming constant $\delta \in [0,1/2)$, which we defined formally in equation (2.8). This robustification of the Sliced Wasserstein distance compares distributions up to a 2δ fraction of their probability mass, thereby generalizing the one-dimensional trimmed Wasserstein distance introduced by Munk and Czado (1998). One of the aims of this chapter is to derive confidence intervals for the trimmed Sliced Wasserstein distance which make either no assumptions or mild moment assumptions on the unknown distributions P and Q. Specifically, given a level $\alpha \in (0,1)$ and i.i.d. samples $X_1,\dots,X_n \sim P$ and $Y_1,\dots,Y_m \sim Q$, we derive confidence sets $C_{nm} \subseteq \mathbb{R}$ such that

$$\inf_{P,Q} \mathbb{P}\big(\mathrm{SW}_{r,\delta}(P,Q) \in C_{nm}\big) \geq 1-\alpha, \tag{2.2}$$

where the infimum is over a suitable family of distributions P, Q.

One of the main reasons that the Wasserstein distance has found many applications is the fact that it is a useful notion of distance under weak assumptions. Unlike the Total Variation, Hellinger, Kullback-Leibler and other divergences, the Wasserstein distance between a pair of distributions can be estimated from samples (optimally) under mild assumptions without requiring any smoothing. However, existing results on *inference* for the univariate Wasserstein distance (Munk and Czado, 1998; Freitag, Munk, and Vogt, 2003; Freitag, Czado, and Munk, 2007; Freitag and Munk, 2005) typically require strong smoothness assumptions and suggest different inferential procedures when $P=Q$ as compared to when $P \neq Q$. In contrast, we construct various assumption-light confidence intervals C_{nm} which have finite-sample validity under weak moment assumptions.

The confidence intervals we construct are adaptive to the regularity of the distributions P and Q, as measured by a functional $\mathrm{SJ}_{r,\delta}$ introduced formally in Section 2.3.1 (equation (2.14)). The magnitude of $\mathrm{SJ}_{r,\delta}(P)$ is largely controlled by the tails of P and by whether its one-dimensional projections have connected support. The one-dimensional counterpart of this functional was identified in the work of Bobkov and Ledoux (2019) who showed that when this functional is finite, the empirical measure of $X_1,\dots,X_n$ converges to P under the Wasserstein distance at the fast rate of $O(1/\sqrt{n})$, assuming $d=1$. On the other hand, when this functional is infinite Bobkov and Ledoux (2019) showed that this convergence happens at a slower rate of $O((1/n)^{1/2r})$. Our work shows that the role of the $\mathrm{SJ}_{r,\delta}$ functional in inference is more nuanced. When $\mathrm{SJ}_{r,\delta}(P)$ and $\mathrm{SJ}_{r,\delta}(Q)$ are finite, our confidence intervals have length scaling at the fast rate of $O(1/\sqrt{n \wedge m})$, mirroring the rates of convergence in the work of Bobkov and Ledoux (2019). On the other hand, when these values are infinite, a dichotomy arises: in full generality, when $\mathrm{SW}_{r,\delta}(P,Q)$ is allowed to take arbitrary (small) values uncertainty quantification is difficult and our intervals can have lengths scaling as $O((1/n \wedge m)^{1/2r})$ in the worst case. However, we find, somewhat surprisingly, even when the $\mathrm{SJ}_{r,\delta}$ functional is infinite, accurate $O(1/\sqrt{n \wedge m})$-inference is possible so long as $\mathrm{SW}_{r,\delta}(P,Q)$ is bounded away from 0. We emphasize that the intervals we construct are *adaptive,* i.e. they have small lengths under appropriate conditions on the $\mathrm{SJ}_{r,\delta}$ functional and $\mathrm{SW}_{r,\delta}(P,Q)$, without needing the statistician to specify or have knowledge of these quantities. We also show that our confidence intervals have minimax optimal length over classes of distributions with varying magnitudes of $\mathrm{SJ}_{r,\delta}(P)$.

To complement our results on confidence intervals for the Sliced Wasserstein distance we also consider the problem of estimating the Sliced Wasserstein distance between two distributions, given samples from each of them. We provide minimax upper and lower bounds for this problem as well. Indeed, our minimax lower bounds for confidence interval length are derived directly from minimax lower bounds for estimating the Sliced Wasserstein distance by noting that the minimax length of a confidence interval is bounded from below by the corresponding minimax estimation rate.

We illustrate the practical significance of our methodology via an application to likelihood-free inference (Sisson, Fan, and Beaumont, 2018), in which a parametrized stochastic simulator for the data-generating process is available, but its underlying distribution is intractable. Here, our goal is to construct confidence intervals for unknown parameters of the simulator, on the basis of minimizing its Sliced Wasserstein distance from an observed sample. Distributional assumptions such as those made in past work on inference for the one-dimensional Wasserstein distance (Munk and Czado, 1998; Freitag, Munk, and Vogt, 2003; Freitag, Czado, and Munk, 2007; Freitag and Munk, 2005) are typically unverifiable in such applications.

Our Contributions. We summarize the contributions of this chapter as follows.

- We define the δ-trimmed Sliced Wasserstein distance $\mathrm{SW}_{r,\delta}$, and the functional $\mathrm{SJ}_{r,\delta}$, generalizing the functional J_r of Bobkov and Ledoux (2019). We show that the finiteness of $\mathrm{SJ}_{r,\delta}(P)$ is a sufficient condition for the empirical measure to estimate P at the parametric rate under the trimmed Sliced Wasserstein distance, and we prove corresponding minimax lower bounds. We also derive minimax rates of estimating the Sliced Wasserstein distance between two distributions, both in the trimmed and untrimmed settings. These rates are sensitive to the magnitude of the $\mathrm{SJ}_{r,\delta}$ functional.
- We propose two-sample confidence intervals for $\mathrm{SW}_{r,\delta}(P,Q)$ which have finite-sample coverage under minimal moment assumptions. We bound the length of our confidence intervals, showing that they are adaptive both to the magnitude of $\mathrm{SJ}_{r,\delta}(P), \mathrm{SJ}_{r,\delta}(Q)$ and to whether or not $P = Q$. These lengths achieve the minimax rate of estimating the Sliced Wasserstein distance, up to polylogarithmic factors.
- We further contrast our finite-sample confidence intervals with asymptotic methods. In particular, under certain regularity conditions, we derive limit laws and show that the bootstrap is consistent in estimating the distribution of the empirical r-Sliced Wasserstein distance for all $r > 1$, whenever $P \neq Q$. We then show how this last assumption may be removed by combining the strengths of our finite-sample intervals and the bootstrap.
- We illustrate our theoretical findings with a simulation study and an application to likelihood-free inference.

2.2 Background and Related Work

In this section, we first provide some further background on the (univariate) Wasserstein distance and its sliced counterpart before turning our attention to a detailed discussion of related work.

2.2.1 The Univariate Wasserstein Distance

Throughout this chapter, we will make use of the well-known fact that the univariate r-Wasserstein distance can be expressed as the L^r norm of the quantile functions of the probability distributions to be compared. Concretely, let $\mathcal{X} \subseteq \mathbb{R}$ and $P, Q \in \mathcal{P}_r(\mathcal{X})$. Let F, G denote the cumulative distribution functions (CDFs) of P and Q, and denote their respective quantile functions by F^{-1} and G^{-1}, where

$$F^{-1}(u) = \inf\{x \in \mathbb{R} : F(x) \geq u\}, \quad \text{for all } u \in [0,1].$$

We extend F^{-1} to be defined over the entire real line under the convention $F^{-1}(u) = \inf(\mathcal{X})$ for all $u < 0$ and $F^{-1}(u) = \sup(\mathcal{X})$ for all $u > 1$, and similarly for G^{-1}. Then, the one-dimensional Wasserstein distance admits the following closed form (Bobkov and Ledoux, 2019):

$$W_r(P,Q) = \left(\int_0^1 \left|F^{-1}(u) - G^{-1}(u)\right|^r du\right)^{1/r}. \tag{2.3}$$

We will also make use of the ∞-Wasserstein distance: when $\mathcal{X}$ is a bounded set, the limit

$$W_\infty(P,Q) := \lim_{r\to\infty} W_r(P,Q) = \sup_{0\leq u\leq 1} \left|F^{-1}(u) - G^{-1}(u)\right| \tag{2.4}$$

exists, and defines a new metric W_∞ on $\mathcal{P}(\mathcal{X})$. The relationship $W_r(P,Q) \leq W_\infty(P,Q)$ shows that W_∞ is a stronger metric than W_r for any $r \geq 1$. In fact, it is strictly stronger: for instance, $W_r(\delta_0, (1-\epsilon)\delta_0 + \epsilon\delta_1) \to 0$ as $\epsilon \to 0$ if and only if r is finite. In contrast, the metrics W_r induce the same (weak) topology for all finite r, when $\mathrm{diam}(\mathcal{X}) < \infty$ (Villani, 2003).

The One-Dimensional Trimmed Wasserstein Distance. Given distributions $P, Q \in \mathcal{P}(\mathbb{R})$ and a trimming constant $\delta \in [0, 1/2)$, Munk and Czado (1998) define the δ-trimmed Wasserstein distance (up to rescaling) by

$$W_{r,\delta}(P,Q) = \left(\frac{1}{1-2\delta}\int_\delta^{1-\delta} \left|F^{-1}(u) - G^{-1}(u)\right|^r du\right)^{\frac{1}{r}}. \tag{2.5}$$

When $\delta = 0$, $W_{r,\delta}$ reduces to the original Wasserstein distance W_r, and when $\delta > 0$, $W_{r,\delta}$ compares the distributions P and Q up to a 2δ fraction of their tail mass. Specifically, let P^δ denote the distribution with CDF $F^\delta(x) = (F(x) - \delta)/(1-2\delta)$, for all $F^{-1}(\delta) \leq x \leq F^{-1}(1-\delta)$, and similarly for Q^δ. Then, Álvarez-Esteban et al. (2008) note that $W_{r,\delta}(P,Q) = W_r(P^\delta, Q^\delta)$.

In addition, we define the trimmed ∞-Wasserstein distance by

$$W_{\infty,\delta}(P,Q) := \lim_{r\to\infty} W_{r,\delta}(P,Q) = \sup_{\delta\leq u\leq 1-\delta} \left|F^{-1}(u) - G^{-1}(u)\right|.$$

2.2.2 The Sliced Wasserstein Distance

Recall the definition of the Sliced Wasserstein distance given in equation (2.1). In view of equation (2.3), the Sliced Wasserstein distance admits the following closed form, for any $P, Q \in \mathcal{P}_r(\mathbb{R}^d)$

$$\mathrm{SW}_r(P,Q) = \left(\int_{\mathbb{S}^{d-1}} \int_0^1 \left| F_\theta^{-1}(u) - G_\theta^{-1}(u) \right|^r du d\mu(\theta) \right)^{\frac{1}{r}}, \tag{2.6}$$

where F_θ^{-1} and G_θ^{-1} are the respective quantile functions of P_θ and Q_θ. Both integrals of the above expression can be approximated via Monte Carlo sampling from $\mathbb{S}^{d-1}$ and from the unit interval $[0,1]$. This fact makes the computation of the Sliced Wasserstein distance significantly simpler than that of the Wasserstein distance. Moreover, the Sliced Wasserstein distance retains some of the qualitative behaviour of the Wasserstein distance, at least for compactly-supported distributions. Indeed, Bonnotte (2013) showed that for any distributions $P, Q \in \mathcal{P}(\{x \in \mathbb{R}^d : \|x\| \leq M\})$, where $M > 0$, we have

$$\mathrm{SW}_r^r(P,Q) \leq c_{d,r} W_r^r(P,Q) \leq C_{d,r} M^{r-1/(d+1)} \mathrm{SW}_r^{1/(d+1)}(P,Q), \tag{2.7}$$

where $C_{d,r} > 0$ is a constant depending on d and r, but not depending on M, and $c_{d,r} = \frac{1}{d} \int_{\mathbb{S}^{d-1}} \|\theta\|_r^r \, d\mu(\theta)$, which is bounded above by $1/d$ whenever $r \geq 2$. In particular, it follows that the metrics W_r and SW_r are topologically equivalent over $\mathcal{P}(\mathcal{X})$ when $\mathrm{diam}(\mathcal{X}) < \infty$. As we shall see, however, the statistical behaviour of the Wasserstein and Sliced Wasserstein distances can differ dramatically for large dimensions d.

Though the original Sliced Wasserstein distance of Rabin et al. (2011) was defined in terms of the uniform distribution μ over $\mathbb{S}^{d-1}$, a straightforward adaptation of Proposition 5.12 of Bonnotte (2013) shows that SW_r remains a metric over $\mathcal{P}(\mathbb{R}^d)$ when μ is replaced by any probability measure which is absolutely continuous with respect to the Hausdorff measure on $\mathbb{S}^{d-1}$. We allow μ to be any such measure throughout the sequel.

The Trimmed Sliced Wasserstein Distance. In analogy to the trimmed Wasserstein distance in equation (2.5), we further define

$$\mathrm{SW}_{r,\delta}(P,Q) = \left(\frac{1}{1-2\delta} \int_{\mathbb{S}^{d-1}} \int_\delta^{1-\delta} \left| F_\theta^{-1}(u) - G_\theta^{-1}(u) \right|^r du d\mu(\theta) \right)^{\frac{1}{r}}, \tag{2.8}$$

for some $\delta \in [0, 1/2)$. We also define the trimmed ∞-Sliced Wasserstein distance by $\mathrm{SW}_{\infty,\delta}(P,Q) = \int_{\mathbb{S}^{d-1}} W_{\infty,\delta}(P_\theta, Q_\theta) d\mu(\theta)$, and more generally, we write

$$\mathrm{SW}_{\infty,\delta}^{(r)}(P,Q) = \int_{\mathbb{S}^{d-1}} W_{\infty,\delta}^r(P_\theta, Q_\theta) d\mu(\theta).$$

When $\delta > 0$, the trimmed-Sliced Wasserstein distance is well-defined and finite for all distributions $P, Q \in \mathcal{P}(\mathbb{R}^d)$, including those admitting fewer than r moments. Nevertheless, it can be easily seen that $\sup_{P,Q\in\mathcal{P}(\mathbb{R}^d)} \mathrm{SW}_{r,\delta}(P,Q) = \infty$. It will be fruitful in our development to impose moment conditions which ensure that the quantity $\mathrm{SW}_{r,\delta}(P,Q)$ is uniformly bounded—one such condition is given in terms of the class

$$\mathcal{K}_{r,\rho}(b) = \left\{ P \in \mathcal{P}(\mathbb{R}^d) : \int_{\mathbb{S}^{d-1}} \mathbb{E}_{X\sim P}\big[|X^\top\theta|^\rho\big]^{\frac{r}{\rho}} d\mu(\theta) \leq b \right\}, \quad b,\rho,r \geq 1. \tag{2.9}$$

We shall use the special case $\rho = 2$ most often, in which case we drop the subscript ρ and simply write $\mathcal{K}_r(b) := \mathcal{K}_{r,2}(b)$. It follows from Lemma 3 in Section 2.A that $\mathrm{SW}_{r,\delta}(P,Q)$ is uniformly bounded over distributions $P,Q \in \mathcal{K}_r(b)$, by a constant depending only on b, r and δ. Notice that if $\bar{b} = b^{\rho/r}$, then $\mathcal{K}_{r,\rho}(b)$ contains the class

$$\overline{\mathcal{K}}_\rho(\bar{b}) = \left\{ P \in \mathcal{P}_\rho(\mathbb{R}^d) : \mathbb{E}_{X\sim P}\big[\, \|X\|^\rho \,\big] \leq \bar{b} \right\}. \tag{2.10}$$

Finally, we write $\mathcal{K}_{r,\rho} = \bigcup_{b\geq 1} \mathcal{K}_{r,\rho}(b)$, $\mathcal{K}_r = \bigcup_{b\geq 1} \mathcal{K}_r(b)$ and $\overline{\mathcal{K}}_\rho = \bigcup_{\bar{b}\geq 1} \overline{\mathcal{K}}_\rho(\bar{b})$.

2.2.3 Related Work

We are unaware of any other work regarding statistical inference for the Sliced Wasserstein distance, except in the special case $d = 1$ when it coincides with the one-dimensional Wasserstein distance. In this case, Munk and Czado (1998) study limiting distributions of the empirical (plug-in) Wasserstein distance estimator, and Freitag, Munk, and Vogt (2003); Freitag, Czado, and Munk (2007); Freitag and Munk (2005) establish sufficient conditions for the validity of the bootstrap in estimating the distribution of the empirical second-order trimmed Wasserstein distance. While these results are very useful, they assume that (i) P and Q are absolutely continuous, (ii) with densities supported on connected sets, and (iii) require different inferential procedures at the classical null ($P = Q$) and away from the null ($P \neq Q$). In contrast, the confidence intervals derived in the present chapter are valid under either no assumptions or mild moment assumptions on P and Q, and are applied more generally to the Sliced Wasserstein distance in arbitrary dimension. Though our methodology is assumption-light, our confidence intervals are adaptive to (iii), and assumptions (i) and (ii) are closely related to the finiteness of $\mathrm{SJ}_{r,\delta}(P), \mathrm{SJ}_{r,\delta}(Q)$, to which our confidence intervals are also adaptive.

The Sliced Wasserstein distance is one of many modifications of the Wasserstein distance based on low-dimensional projections. We mention here the Generalized Sliced (Kolouri et al., 2019), Tree-Sliced (Le et al., 2019), max-Sliced (Deshpande et al., 2019), Subspace Robust (Paty and Cuturi, 2019; Niles-Weed and Rigollet, 2022), and Distributional Sliced (Nguyen et al., 2020) Wasserstein distances. It is also possible to define various other interesting distances by slicing (averaging along univariate projections—see Kim, Balakrishnan, and Wasserman (2020)).

Beyond the aforementioned inferential results for the one-dimensional Wasserstein distance, statistical inference for Wasserstein distances over finite or countable spaces has been studied by Sommerfeld and Munk (2018); Tameling, Sommerfeld, and Munk (2019); Klatt, Tameling, and Munk (2020); Klatt, Munk, and Zemel (2022). For distributions with multidimensional support, Rippl, Munk, and Sturm (2016) consider the situation where P and Q only differ by a location-scale transformation. Imaizumi, Ota, and Hamaguchi (2022) study the validity of the multiplier bootstrap for estimating the distribution of the plug-in estimator of an approximation of the Wasserstein distance. Central Limit Theorems for empirical Wasserstein distances in general dimension have been established by del Barrio and Loubes (2019); del Barrio, González-Sanz, and Loubes (2021), but with unknown centering constants which are a barrier to using these results for statistical inference.

Rates of convergence for the problem of estimating a distribution under the Wasserstein distance (Dudley (1969); Boissard and Le Gouic (2014); Fournier and Guillin (2015); Bobkov and Ledoux (2019); Weed and Bach (2019); Singh and Póczos (2019); Lei (2020), and references therein) have received significantly more attention than the problem of estimating the Wasserstein distance, the latter being more closely related to our work. Minimax rates of estimating the Wasserstein distance between two distributions have been established by Niles-Weed and Rigollet (2022), as well as by Liang (2019) when $r = 1$; we will also return to this topic in the following chapter. In the special case $d = 1$, where the Sliced Wasserstein distance coincides with the Wasserstein distance, our results refine those of Niles-Weed and Rigollet (2022) by showing that faster rates can be achieved depending on the finiteness of the $\mathrm{SJ}_{r,\delta}$ functional, and on the magnitude of $\mathrm{SW}_{r,\delta}(P, Q)$.

Likelihood-free inference methodology with respect to the Wasserstein and Sliced Wasserstein distances has recently been developed by Bernton et al. (2019) and Nadjahi et al. (2020). In contrast to these methods, both of which employ approximate Bayesian computation, our work provides frequentist coverage guarantees under minimal assumptions.

2.3 Estimating the Sliced Wasserstein Distance

The goal of this section is to bound the minimax risk of estimating the Sliced Wasserstein distance between two distributions, that is

$$\mathcal{R}_{nm} \equiv \mathcal{R}_{nm}(\mathcal{O}; r) = \inf_{\widehat{S}_{nm}} \sup_{(P,Q)\in\mathcal{O}} \mathbb{E}_{P^{\otimes n}\otimes Q^{\otimes m}} \big|\widehat{S}_{nm} - \mathrm{SW}_{r,\delta}(P, Q)\big|, \tag{2.11}$$

where the infimum is over all estimators $\widehat{S}_{nm}$ of the Sliced Wasserstein distance based on a sample of size n from P and a sample of size m from Q, and $\mathcal{O} \subseteq \mathcal{P}(\mathbb{R}^d) \times \mathcal{P}(\mathbb{R}^d)$ is a collection of pairs of distributions. Our motivation for studying this quantity is the observation that $\mathcal{R}_{nm}$ lower bounds the minimax length of a confidence interval for the Sliced Wasserstein distance. We construct confidence intervals with matching length in Section 2.4.

The estimation problem in equation (2.11) is related to, but distinct from, the problem of estimating a distribution under the Sliced Wasserstein distance. The minimax risk associated with this problem is given by

$$\mathcal{M}_n \equiv \mathcal{M}_n(\mathcal{J}; r) = \inf_{\widehat{P}_n} \sup_{P\in\mathcal{J}} \mathbb{E}_{P^{\otimes n}} \Big\{\mathrm{SW}_{r,\delta}(\widehat{P}_n, P)\Big\}, \tag{2.12}$$

where the infimum is over all estimators $\widehat{P}_n$ of Borel probability distributions P, based on a sample of size n from P, and $\mathcal{J} \subseteq \mathcal{P}(\mathbb{R}^d)$. Problems (2.11) and (2.12) are related as follows:

Given estimators $\widehat{P}_n$ and $\widehat{Q}_m$ for two distributions $P, Q \in \mathcal{P}(\mathbb{R}^d)$, which are minimax optimal in the sense of equation (2.12), we have, by the triangle inequality,

$$\begin{aligned} \mathcal{R}_{nm}(\mathcal{O}; r) &\lesssim \mathbb{E}\big|\mathrm{SW}_{r,\delta}(\widehat{P}_n, \widehat{Q}_m) - \mathrm{SW}_{r,\delta}(P, Q)\big| \\ &\leq \mathbb{E}\mathrm{SW}_{r,\delta}(\widehat{P}_n, P) + \mathbb{E}\mathrm{SW}_{r,\delta}(\widehat{Q}_m, Q) \lesssim \mathcal{M}_{n\wedge m}(\mathcal{J}; r), \end{aligned} \tag{2.13}$$

for suitable families $\mathcal{J}$ and $\mathcal{O}$ (typically $\mathcal{O} \subseteq \mathcal{J} \times \mathcal{J}$). Inequality (2.13) implies that estimating a distribution under $\mathrm{SW}_{r,\delta}$ is a more challenging problem, statistically, than that of estimating the Sliced Wasserstein distance between two distributions. It is unclear, however, whether the rate $\mathcal{M}_{n\wedge m}$ is a tight upper bound on $\mathcal{R}_{nm}$, or whether the latter can be further reduced. For the Wasserstein distance W_r in general dimension, Liang (2019) and Niles-Weed and Rigollet (2022) showed that there is no gap between these minimax risks (ignoring polylogarithmic factors) for compactly supported distributions.

Let us now briefly summarize the main results of this section. We bound $\mathcal{M}_n$ and $\mathcal{R}_{nm}$, and show that there can be a large gap between these minimax risks when the pairs of distributions in $\mathcal{O}$ are appropriately separated. In the special case $d = 1$, $\mathrm{SW}_{r,\delta}$ reduces to the (trimmed) Wasserstein distance, and our results imply faster rates than those of Liang (2019) and Niles-Weed and Rigollet (2022), for estimating the Wasserstein distance between distributions bounded away from each other. Furthermore, in contrast to the minimax risk for estimating the Wasserstein distance and estimating under the Wasserstein distance, the minimax risks we obtain for the Sliced Wasserstein distance when $d > 1$ are dimension-free.

Though our primary interest is in $\mathcal{R}_{nm}$ (due to its direct connection to confidence intervals) we begin by studying $\mathcal{M}_n$ to motivate our choices of families $\mathcal{O}$. Inspired by Bobkov and Ledoux (2019), in Section 2.3.1 we define a functional $\mathrm{SJ}_{r,\delta}$, whose magnitude is related to the regularity of the supports of P and Q, and whose finiteness implies improved rates of decay for $\mathcal{M}_n$. We then study the minimax risk $\mathcal{R}_{nm}$ over various families $\mathcal{O}$ in Section 2.3.2.

2.3.1 Minimax Estimation under the Sliced Wasserstein Distance

Let $\delta \in [0, 1/2)$, $P \in \mathcal{P}(\mathbb{R}^d)$, and let $X_1, \dots, X_n \sim P$ be an i.i.d. sample. Let $P_n = \frac{1}{n}\sum_{i=1}^n \delta_{X_i}$ denote the corresponding empirical measure. The goal of this section is to characterize the rates of convergence of P_n to the distribution P under the (trimmed) Sliced Wasserstein distance, extending the comprehensive treatment by Bobkov and Ledoux (2019). We then provide corresponding minimax lower bounds on $\mathcal{M}_n$.

For any $\theta \in \mathbb{S}^{d-1}$, let p_θ denote the density of the absolutely continuous component in the Lebesgue decomposition of the measure $P_\theta = \pi_{\theta\#}P$. Define the functional

$$\mathrm{SJ}_{r,\delta}(P) = \frac{1}{1-2\delta}\int_{\mathbb{S}^{d-1}}\int_{\delta}^{1-\delta}\left(\frac{\sqrt{u(1-u)}}{p_\theta(F_\theta^{-1}(u))}\right)^r du d\mu(\theta), \tag{2.14}$$

with the convention that $0/0 = 0$. When $d = 1$, we write $\mathrm{J}_{r,\delta}$ instead of $\mathrm{SJ}_{r,\delta}$, and in the untrimmed case $\delta = 0$, we omit the subscript δ and write SJ_r or J_r. When $d = 1$ and $\delta = 0$, Bobkov and Ledoux (2019) prove that the finiteness of $\mathrm{J}_r(P)$ is a necessary and sufficient condition for $\mathbb{E}\big[W_r(P_n,P)\big]$ to decay at the parametric rate $n^{-1/2}$. The magnitude of J_r is thus closely related to the convergence behaviour of empirical measures under one-dimensional Wasserstein distances, and we show below that the same is true for the $\mathrm{SJ}_{r,\delta}$ functional with respect to trimmed Sliced Wasserstein distances, using distinct proof techniques.

It can be seen that a necessary condition for the finiteness of $\mathrm{SJ}_{r,\delta}(P)$ is that for μ-almost every $\theta \in \mathbb{S}^{d-1}$, the density p_θ is supported on a (possibly infinite) interval. When δ vanishes, the value of $\mathrm{SJ}_{r,\delta}(P)$ also depends on the tail behaviour of P and the value of r. For example, if $P = N(0, I_d)$ is the standard Gaussian distribution, it can be shown that $\mathrm{SJ}_{r,\delta}(P) < \infty$ whenever $\delta > 0$, whereas for $\delta = 0$, $\mathrm{SJ}_r(P) < \infty$ if and only if $1 \leq r < 2$ by a similar argument as Bobkov and Ledoux (2019, p. 46). On the other hand, if $P = \frac{1}{2}U(0,\Delta_1) + \frac{1}{2}U(\Delta_2, 1)$, for some $0 < \Delta_1 \leq \Delta_2 < 1$, where $U(a,b)$ denotes the uniform distribution on the interval $(a,b) \subseteq \mathbb{R}$, one has $\mathrm{SJ}_{r,\delta}(P) < \infty$ if and only if $\Delta_1 = \Delta_2$, for every $\delta \in [0, 1/2)$.

We now provide two upper bounds on $\mathbb{E}\big[\mathrm{SW}_{r,\delta}(P_n,P)\big]$, which are effective when $\mathrm{SJ}_{r,\delta}(P) < \infty$ and $\mathrm{SJ}_{r,\delta}(P) = \infty$ respectively. Recall the class $\mathcal{K}_r(b)$ from equation (2.9).

Proposition 2. Let $b, r \geq 1$ and $\delta \in (0, 1/2)$. Assume $P \in \mathcal{K}_r(b)$, and that $\delta \geq 2(r+2)/n$.

(i) There exist constants $c_r, c_r' > 0$ depending only on r such that

$$\mathbb{E}\big[\mathrm{SW}_{r,\delta}(P_n,P)\big] \leq c_r \mathrm{SJ}_{r,\frac{\delta}{2}}^{\frac{1}{r}}(P)\sqrt{\frac{\log n}{n}} + \frac{c_r(bne^{-c_r' n\delta})^{\frac{1}{r}}}{\sqrt{\delta}}. \tag{2.15}$$

(ii) There exists a constant $k_r > 0$ depending only on r such that

$$\mathbb{E}\big[\mathrm{SW}_{r,\delta}(P_n,P)\big] \leq \frac{k_r}{\sqrt{\delta}}\left(\frac{b}{1-2\delta}\right)^{\frac{1}{r}} n^{-1/2r}. \tag{2.16}$$

Proposition 2 provides two upper bounds on the rate of convergence of the empirical measure under $\mathrm{SW}_{r,\delta}$, which are closely related to those of Theorem 5.3 and Theorem 7.16 of Bobkov and Ledoux (2019) for the one-dimensional untrimmed Wasserstein distance. Bobkov and Ledoux (2019) established these results by hinging upon a representation of the empirical one-dimensional Wasserstein distance in terms of the so-called mean square beta distribution, coupled with Poincaré-type inequalities for such measures. While extensions of these techniques to the *untrimmed* Sliced Wasserstein distance are straightforward, and will be stated for completeness in Section 2.3.3, it was unclear to us whether they may be adapted to the *trimmed* setting $\delta > 0$. Our proof of Proposition 2 is instead based on uniform concentration inequalities for the empirical quantile process—an approach which we now describe, and which foreshadows the construction of our confidence intervals in Section 2.4.

Proposition 2(i) is proved using a uniform bound for self-normalized empirical processes, known as the relative Vapnik-Chervonenkis (VC) inequality (Vapnik, 2013; Bousquet, Boucheron, and Lugosi, 2003), which will be further described in Example 2 below. This result shows that the empirical measure converges at the parametric rate under $\mathrm{SW}_{r,\delta}$, up to a polylogarithmic factor, provided $\mathrm{SJ}_{r,\delta/2}(P)$ is bounded, and provided, for instance, that the trimming constant

δ does not vanish at a rate faster than $n^{-\beta}$ for some $\beta \in (0,1)$. We emphasize that this convergence is uniform in P—for instance, one has that for all $s \geq 1$,

$$\sup_{\substack{P \in \mathcal{K}_r(b) \\ \mathrm{SJ}_{r,\delta/2}(P) \leq s}} \mathbb{E}\big[\mathrm{SW}_{r,\delta}(P_n, P)\big] \underset{b,r}{\lesssim} s^{\frac{1}{r}} \sqrt{\frac{\log n}{n}}, \text{ when } \delta \asymp n^{-\beta}, \text{for some } \beta \in (0,1).$$

In fact, the above bound continues to hold when $s \leq 1$—a regime relevant for distributions with vanishing variances—so long as $s^{1/r}\sqrt{\log n / n}$ remains greater than the second term in equation (2.15). Finally, we note that the polylogarithmic factor in the above display arises from the relative VC inequality. Example 2.7 of Giné and Koltchinskii (2006) suggests that this factor may be improved to $\sqrt{\log \log n}$, but not further if $\delta \gtrsim 1/n$. We do not know whether this factor can be omitted if stronger conditions are placed on δ while allowing it to vanish.

Proposition 2(ii) is primarily of interest for distributions such that $\mathrm{SJ}_{r,\delta}(P) = \infty$, and is proved using the Dvoretzky-Kiefer-Wolfowitz inequality (Dvoretzky, Kiefer, and Wolfowitz, 1956; Massart, 1990). This result shows that the empirical measure converges to P at the nonparametric rate $n^{-1/2r}$ under no assumptions on P apart from the mild moment assumption $P \in \mathcal{K}_r(b)$. In contrast to Proposition 2(i), however, this result suffers from a markedly worse dependence on δ. Indeed, the resulting rate of convergence deteriorates as soon as $\delta = o(1)$, and we conjecture that this behaviour is necessary under the stated assumptions.

The rates in Proposition 2 do not depend on the dimension d, contrasting generic rates of convergence of the empirical measure under the Wasserstein distance. For instance, if P is supported on a bounded set in $\mathbb{R}^d$, Lei (2020) (see also Fournier and Guillin (2015), Weed and Bach (2019)) shows that $\mathbb{E}[W_r(P_n, P)] \lesssim n^{-1/d}$ whenever $d > 2r$, and this rate is known to be tight (Dudley, 1969; Singh and Póczos, 2019). Thus, estimating a distribution in the Sliced Wasserstein distance does not suffer from the curse of dimensionality despite metrizing the same topology on $\mathcal{P}(\mathbb{R}^d)$—see equation (2.7).

For completeness, we close this subsection by stating a lower bound on the minimax risk $\mathcal{M}_n$ in equation (2.12). In view of Proposition 2, it is natural to carry out our analysis over the class of distributions

$$\mathcal{J}(s) = \big\{P \in \mathcal{K}_r(b) : \mathrm{SJ}_{r,\delta}(P) \leq s\big\}, \quad s \in [0, \infty].$$

Proposition 3. Let $b, r \geq 1$ and $\delta \in (0, 1/2)$. Then, there exist constants $C_1, C_2 > 0$, possibly depending on b, r, δ, such that for all $s > 0$ satisfying $b \geq (2s)^{1/r}$,

$$\mathcal{M}_n(\mathcal{J}(s); r) \geq C_1 s^{\frac{1}{r}} n^{-1/2}, \quad \text{and} \quad \mathcal{M}_n(\mathcal{J}(\infty); r) \geq C_2 n^{-1/2r}.$$

Proposition 3 implies that the rates achieved by the empirical measure in Proposition 2 are minimax optimal over the classes considered above, up to polylogarithmic factors. The proof of this result will follow as a special case of our bounds on the minimax risk $\mathcal{R}_{nm}$, to which we turn our attention next.

2.3.2 Minimax Estimation of the Sliced Wasserstein Distance

In this section, we bound the minimax risk $\mathcal{R}_{nm}$ defined in equation (2.11). We begin by providing upper bounds on the estimation error of the empirical Sliced Wasserstein distance, $\mathrm{SW}_{r,\delta}(P_n, Q_m)$. Recall that, P_n and Q_m denote the empirical measures of i.i.d. samples $X_1, \dots, X_n \sim P$ and $Y_1, \dots, Y_m \sim Q$, respectively.

Proposition 4. Let $b, r \geq 1$ and $\delta \in (0, 1/2)$. Assume $P, Q \in \mathcal{K}_r(b)$, and that $\delta \geq 2(r+2)/(n \wedge m)$.

(i) There exists a constant $c_r > 0$, possibly depending on r, such that

$$\begin{aligned}
&\mathbb{E}\big|\mathrm{SW}_{r,\delta}(P_n, Q_m) - \mathrm{SW}_{r,\delta}(P, Q)\big| \\
&\quad \underset{r}{\lesssim} \left(\frac{b}{1-2\delta}\right)^{\frac{1}{r}} \frac{n^{-\frac{1}{2r}}}{\sqrt{\delta}} \wedge \left(\mathrm{SJ}^{\frac{1}{r}}_{r,\frac{\delta}{2}}(P) \sqrt{\frac{\log n}{n}} + \frac{(bne^{-c_r n\delta})^{\frac{1}{r}}}{\sqrt{\delta}} \right) \\
&\quad + \left(\frac{b}{1-2\delta}\right)^{\frac{1}{r}} \frac{m^{-\frac{1}{2r}}}{\sqrt{\delta}} \wedge \left(\mathrm{SJ}^{\frac{1}{r}}_{r,\frac{\delta}{2}}(Q) \sqrt{\frac{\log m}{m}} + \frac{(bme^{-c_r m\delta})^{\frac{1}{r}}}{\sqrt{\delta}} \right).
\end{aligned}$$

(ii) Suppose $\mathrm{SW}_{r,\delta}(P, Q) \geq \Gamma$, for some real number $\Gamma > 0$. Then,

$$\mathbb{E}\big|\mathrm{SW}_{r,\delta}(P_n, Q_m) - \mathrm{SW}_{r,\delta}(P, Q)\big| \underset{\Gamma, r}{\lesssim} \frac{b}{\delta^{r/2}(1-2\delta)} \left(n^{-\frac{1}{2}} + m^{-\frac{1}{2}}\right).$$

Proposition 4(i) is an immediate consequence of inequality (2.13), which implies that the rate of estimating the Sliced Wasserstein distance with the plug-in estimator $\mathrm{SW}_{r,\delta}(P_n, Q_m)$ is no worse than the rate of convergence of the empirical measures under $\mathrm{SW}_{r,\delta}$ given in Proposition 2. In particular, these results show that the parametric rate for estimating $\mathrm{SW}_{r,\delta}$ is achievable for distributions satisfying $\mathrm{SJ}_{r,\delta/2}(P), \mathrm{SJ}_{r,\delta/2}(Q) < \infty$, while the rate $n^{-1/2r} + m^{-1/2r}$ is otherwise achievable. On the other hand, Proposition 4(ii) implies that the parametric rate of estimating $\mathrm{SW}_{r,\delta}(P, Q)$ is *always* achievable when P and Q are bounded away from each other under $\mathrm{SW}_{r,\delta}$. This fast rate of convergence is obtained irrespective of the values of $\mathrm{SJ}_{r,\delta}(P)$ and $\mathrm{SJ}_{r,\delta}(Q)$. Discrepancies between rates of convergence at the null ($P = Q$) and away from the null ($P \neq Q$) have previously been noted by Sommerfeld and Munk (2018) for Wasserstein distances over finite spaces—indeed, their rates match those of Proposition 4 when $\mathrm{SJ}_{r,\delta/2}(P), \mathrm{SJ}_{r,\delta/2}(Q) = \infty$. Finally, we note that the natural estimator $\mathrm{SW}_{r,\delta}(P_n, Q_m)$ is *adaptive* to the typically unknown quantities $\mathrm{SJ}_{r,\delta/2}(P)$ and $\mathrm{SJ}_{r,\delta}(Q)$, and does not require the statistician to specify if $P = Q$ or $P \neq Q$. Instead, the estimator adapts and yields favorable rates in favorable situations—when either the $\mathrm{SJ}_{r,\delta/2}$ functionals are finite, or when P and Q are sufficiently well-separated.

We now provide corresponding lower bounds on the minimax risk $\mathcal{R}_{nm}$. Inspired by Proposition 4, we define the following collection of pairs of distributions,

$$\mathcal{O}(\Gamma; s_1, s_2) = \Big\{ (P, Q) \in \mathcal{K}_r^2(b) : \mathrm{SJ}_{r,\delta}(P) \leq s_1, \mathrm{SJ}_{r,\delta}(Q) \leq s_2, \mathrm{SW}_{r,\delta}(P, Q) \geq \Gamma \Big\},$$

where $s_1, s_2 \in [0, \infty]$ and $\Gamma \geq 0$. To ensure that the class $\mathcal{O}(\Gamma; s_1, s_2)$ is nonempty, we assume in what follows that $\Gamma^r \leq c_r b$, for some sufficiently small constant $c_r > 0$ depending only on r. With these definitions in place we now state our minimax lower bounds on the risk $\mathcal{R}_{nm}$.

Theorem 5. *Let $b, r \geq 1$ and $\delta \in (0, 1/2)$. Fix $s > 0$, and assume $b \geq (2s)^{1/r}$.*

(i) There exists a constant $C_1 > 0$, possibly depending on δ, r, b, such that for any $s_1, s_2 \in [0, \infty]$,

$$\mathcal{R}_{nm}(\mathcal{O}(0; s_1, s_2); r) \geq C_1 \begin{cases} n^{-\frac{1}{2r}} + m^{-\frac{1}{2r}}, & s_1 = s_2 = \infty \\ \frac{s_1^{\frac{1}{r}}}{\sqrt{n}} + \frac{s_2^{\frac{1}{r}}}{\sqrt{m}}, & s_1 \vee s_2 \leq s. \end{cases}$$

(ii) For any $\Gamma > 0$ such that $\Gamma^r \leq c_r b$, there exists a constant $C_2 > 0$ possibly depending on δ, r, b, Γ such that

$$\mathcal{R}_{nm}(\mathcal{O}(\Gamma; \infty, \infty); r) \geq C_2 \left(n^{-\frac{1}{2}} + m^{-\frac{1}{2}} \right).$$

Theorem 5 implies that the rates achieved by the empirical Sliced Wasserstein distance $\mathrm{SW}_r(P_n, Q_m)$ in Proposition 4, including their dependence on the $\mathrm{SJ}_{r,\delta}$ functional, are minimax optimal (ignoring polylogarithmic factors). We defer its proof to Section 2.C. This result is proved by a standard information-theoretic technique of constructing pairs of distributions which are statistically indistinguishable but have very different Sliced Wasserstein distances. We then obtain lower bounds via an application of Le Cam's Lemma (see, for instance, Theorem 2.2 of Tsybakov (2008)). Beyond this careful choice of distributions, the bulk of our technical effort lies in tightly bounding the various Sliced Wasserstein distances (see Lemma 10 in Section 2.C).

In Section 2.4, we construct finite-sample confidence intervals for $\mathrm{SW}_{r,\delta}(P, Q)$ whose lengths achieve these same rates of convergence, up to polylogarithmic factors. Before turning to these results, we discuss estimation rates in the untrimmed case when $\delta = 0$.

2.3.3 Minimax Estimation of the Untrimmed Sliced Wasserstein Distance

Though our results below on finite-sample and asymptotic inference will focus on the trimmed Sliced Wasserstein distance, as it is an estimand for which assumption-free inference is possible, we end this section by deriving convergence rates for estimating the untrimmed Sliced Wasserstein distance. In this setting, a straightforward extension of Theorem 5.3 and Theorem 7.16 of Bobkov and Ledoux (2019) already leads to the following untrimmed analogue of Proposition 2, which we state for completeness. We recall that the class $\mathcal{K}_{r,\rho}(b)$ is defined in equation (2.9).

Proposition 5. Let $r \geq 1$ and $s > 0$. Then,

$$\sup_{\substack{P \in \mathcal{P}(\mathbb{R}^d) \\ \mathrm{SJ}_r(P) \leq s}} \mathbb{E}\mathrm{SW}_r(P_n, P) \lesssim_r s^{\frac{1}{r}} n^{-\frac{1}{2}}.$$

Furthermore, for any $\rho > 2r$ and $b > 0$,

$$\sup_{P \in \mathcal{K}_{r,\rho}(b)} \mathbb{E}\mathrm{SW}_r(P_n, P) \lesssim_{b,\rho,r} n^{-\frac{1}{2r}}.$$

Convergence rates for estimating $\mathrm{SW}_r(P,Q)$ immediately follow from Proposition 5. For example, we obtain

$$\sup_{P,Q \in \mathcal{K}_{r,\rho}(b)} \mathbb{E}\big|\mathrm{SW}_r(P_n, Q_m) - \mathrm{SW}_r(P,Q)\big| \lesssim_{b,\rho,r} n^{-\frac{1}{2r}} + m^{-\frac{1}{2r}}.$$

By analogy with Proposition 4(ii), it is natural to ask whether the above convergence rate may be improved to the parametric rate when P and Q are separated in Sliced Wasserstein distance. Such an assertion cannot be deduced from the work of Bobkov and Ledoux (2019), and is the subject of the following main result.

Theorem 6. *For any $r, b \geq 1$, $\Gamma > 0$, and any $\rho > 2r$, it holds that*

$$\sup_{\substack{P,Q \in \mathcal{K}_{r,\rho}(b) \\ \mathrm{SW}_r(P,Q) \geq \Gamma}} \mathbb{E}\big|\mathrm{SW}_r(P_n, Q_m) - \mathrm{SW}_r(P,Q)\big| \lesssim_{\Gamma,\rho,r} b\left(\sqrt{\frac{\log n}{n}} + \sqrt{\frac{\log m}{m}}\right). \tag{2.17}$$

Theorem 6 proves that the parametric rate for estimating the Sliced Wasserstein distance between well-separated distributions continues to hold in the absence of trimming, at the price of a polylogarithmic factor. In fact, our proof shows more generally that the following bound holds without separation conditions on P and Q,

$$\sup_{P,Q \in \mathcal{K}_{r,\rho}(b)} \mathbb{E}\big|\mathrm{SW}_r^r(P_n, Q_m) - \mathrm{SW}_r^r(P,Q)\big| \lesssim_{\rho,r} b\left(\sqrt{\frac{\log n}{n}} + \sqrt{\frac{\log m}{m}}\right). \tag{2.18}$$

The only regularity condition required for these bounds is that $P, Q \in \mathcal{K}_{r,\rho}(b)$, for some $\rho > 2r$. When $d = 1$, this condition is equivalent to assuming that P and Q have finite moments of order ρ, and is otherwise weaker when $d > 1$. The threshold $\rho > 2r$ appears to be nearly sharp, at least for the conclusion of equation (2.18) to hold. One clearly requires $\rho \geq r$, as otherwise $\mathrm{SW}_r(P,Q)$ may be infinite. In the range $r < \rho < 2r$, for the special case $d = 1$ and $P = Q$, Fournier and Guillin (2015) argue that the rate in equation (2.18) cannot be improved beyond $n^{-\frac{\rho - r}{\rho}}$, which is polynomially slower than the parametric rate. While we do not know the sharp rate in this regime when $P \neq Q$, we expect that the parametric rate is not achievable even under this restriction, for $\rho < 2r$.

Theorem 6 is proved using a peeling argument, coupled with a uniform self-normalized concentration inequality for the empirical quantile process, which was already discussed following the statement of Proposition 2. Unlike the latter result, where this inequality allowed us to obtain rates which adapt to the $\mathrm{SJ}_{r,\delta}$ functional, its use here is essential for obtaining a nearly sharp rate without unnecessary moment assumptions, as it allows us to tightly control the behaviour of extremal empirical quantiles. We defer the proof to Section 2.D.

2.4 Finite-Sample Confidence Intervals

2.4.1 Finite-Sample Confidence Intervals in Dimension One

Throughout this subsection, let $r \geq 1$ and $\delta \in [0, 1/2)$ be given, let $P, Q \in \mathcal{P}(\mathbb{R})$ be probability distributions with respective CDFs F, G, and let $X_1, \dots, X_n \sim P$ and $Y_1, \dots, Y_m \sim Q$ be i.i.d. samples. Let $F_n(x) = \frac{1}{n}\sum_{i=1}^n I(X_i \leq x)$ and $G_m(x) = \frac{1}{m}\sum_{j=1}^m I(Y_j \leq x)$ denote their corresponding empirical CDFs, for all $x \in \mathbb{R}$. We derive confidence intervals $C_{nm} \subseteq \mathbb{R}$ for the δ-trimmed Wasserstein distance, with the following non-asymptotic coverage guarantee

$$\inf_{P,Q \in \mathcal{P}(\mathbb{R})} \mathbb{P}\big(W_{r,\delta}(P, Q) \in C_{nm}\big) \geq 1 - \alpha, \tag{2.19}$$

for some pre-specified level $\alpha \in (0, 1)$. Our approach hinges on the fact that the one-dimensional Wasserstein distance may be expressed as the L^r norm of the quantile functions of P and Q (cf. equation (2.3)), suggesting that a confidence interval may be derived via uniform control of the empirical quantile process. Specifically, the starting point for our confidence intervals is a confidence band of the form

$$\inf_{P \in \mathcal{P}(\mathbb{R})} \mathbb{P}\Big(F_n^{-1}\big(\gamma_{\alpha,n}(u)\big) \leq F^{-1}(u) \leq F_n^{-1}\big(\eta_{\alpha,n}(u)\big),\ \forall u \in (0, 1)\Big) \geq 1 - \alpha/2, \tag{2.20}$$

for some sequences of functions $\gamma_{\alpha,n}, \eta_{\alpha,n} : (0, 1) \to \mathbb{R}$. The study of uniform quantile bounds of the form (2.20) is a classical topic (see for instance, the book of Shorack and Wellner (2009)). We discuss two prominent examples that will form the basis of our development.

Example 1. By the Dvoretzky-Kiefer-Wolfowitz (DKW) inequality (Dvoretzky, Kiefer, and Wolfowitz, 1956; Massart, 1990), we have

$$\mathbb{P}\Big(|F_n(x) - F(x)| \leq \beta_n,\ \forall x \in \mathbb{R}\Big) \geq 1 - \frac{\alpha}{2}, \quad \beta_n = \sqrt{\frac{1}{2n}\log(4/\alpha)}. \tag{2.21}$$

Inverting this inequality leads to the choice

$$\gamma_{\alpha,n}(u) = u - \beta_n, \quad \eta_{\alpha,n}(u) = u + \beta_n, \quad u \in (0, 1). \tag{2.22}$$

Example 2. Scale-dependent choices of $\gamma_{\alpha,n}$ and $\eta_{\alpha,n}$ may be obtained via the relative Vapnik-Chervonenkis (VC) inequality (Vapnik, 2013). The latter implies the inequality

$$\mathbb{P}\left(|F_n(x) - F(x)| \leq \nu_{\alpha,n}\sqrt{F_n(x)(1 - F_n(x))},\ \forall x \in \mathbb{R}\right) \geq 1 - \frac{\alpha}{2}, \tag{2.23}$$

where $\nu_{\alpha,n} := \sqrt{\frac{16}{n}\left[\log(16/\alpha) + \log(2n + 1)\right]}$. As shown in Section 2.G.5, inverting inequality (2.23) leads to the following choice, for all $u \in (0, 1)$,

$$\begin{aligned} \gamma_{\alpha,n}(u) &= \frac{2u + \nu_{\alpha,n}^2 - \nu_{\alpha,n}\sqrt{\nu_{\alpha,n}^2 + 4u(1 - u)}}{2(1 + \nu_{\alpha,n}^2)}, \\ \eta_{\alpha,n}(u) &= \frac{2u + \nu_{\alpha,n}^2 + \nu_{\alpha,n}\sqrt{\nu_{\alpha,n}^2 + 4u(1 - u)}}{2(1 + \nu_{\alpha,n}^2)}. \end{aligned} \tag{2.24}$$

Given sequences of functions $\gamma_{\alpha,n}, \eta_{\alpha,n}$ satisfying equation (2.20), one has with probability at least $1-\alpha$, $A_{nm}(u) \leq |F^{-1}(u) - G^{-1}(u)| \leq B_{nm}(u)$ uniformly in $u \in [\delta, 1-\delta]$, where,

$$\begin{aligned} A_{nm}(u) &= \big[F_n^{-1}\big(\gamma_{\alpha,n}(u)\big) - G_m^{-1}\big(\eta_{\alpha,m}(u)\big)\big] \vee \big[G_m^{-1}\big(\gamma_{\alpha,m}(u)\big) - F_n^{-1}\big(\eta_{\alpha,n}(u)\big)\big] \vee 0, \\ B_{nm}(u) &= \big[F_n^{-1}\big(\eta_{\alpha,n}(u)\big) - G_m^{-1}\big(\gamma_{\alpha,m}(u)\big)\big] \vee \big[G_m^{-1}\big(\eta_{\alpha,m}(u)\big) - F_n^{-1}\big(\gamma_{\alpha,n}(u)\big)\big]. \end{aligned}$$

This observation readily leads to the following Proposition.

Proposition 6. Let $\delta \in [0, 1/2)$ and $r \geq 1$. Then, the interval

$$C_{nm} = \left[\left(\frac{1}{1-2\delta}\int_\delta^{1-\delta} A_{nm}^r(u)du\right)^{\frac{1}{r}}, \left(\frac{1}{1-2\delta}\int_\delta^{1-\delta} B_{nm}^r(u)du\right)^{\frac{1}{r}}\right], \tag{2.25}$$

satisfies $\inf\limits_{P,Q\in\mathcal{P}(\mathbb{R})} \mathbb{P}\big(W_{r,\delta}(P,Q) \in C_{nm}\big) \geq 1-\alpha$.

Proposition 6 establishes the finite-sample coverage of the confidence interval C_{nm}, under no assumptions on the distributions P, Q. We emphasize, however, that for distributions P, Q with unbounded support, the interval C_{nm} only has finite length under the following condition.

A1($\delta; \alpha$) We have $\gamma_{\alpha,n\wedge m}(\delta) > 0$ and $\eta_{\alpha,n\wedge m}(1-\delta) < 1$.

If $\gamma_{\alpha,n}, \eta_{\alpha,n}$ are chosen via the DKW inequality (2.22), these inequalities imply the choice $\delta \gtrsim (n\wedge m)^{-1/2}$, while if they are chosen via the relative VC inequality (2.24), one must take $\delta \gtrsim \log(n\wedge m)(n\wedge m)^{-1}$. These choices exclude the untrimmed case $\delta = 0$, for which statistical inference for the Wasserstein distance is not possible without any assumptions on the tail behaviour of P and Q. If explicit bounds on the quantile functions of P and Q are known near the boundary of the unit interval—which is for instance the case when an upper bound on the moments of P and Q is known—these may be used to replace the confidence band $[F_n^{-1}(\gamma_{\alpha,n}(u)), F_n^{-1}(\eta_{\alpha,n}(u))]$ by one of finite length for values of $u \in [0,1]$ satisfying $\gamma_{\alpha,n}(u) < 0$ and $\eta_{\epsilon,n}(u) > 1$. Doing so would lead to a confidence interval C_{nm} of finite length for the untrimmed Wasserstein distance. Since our goal is assumption-free inference, however, we do not pursue this avenue here and we therefore assume **A1($\delta; \alpha$)** holds throughout the sequel.

2.4.2 Finite-Sample Confidence Intervals in General Dimension

We now use Proposition 6 to derive a confidence interval for $\mathrm{SW}_{r,\delta}(P,Q)$, where $P, Q \in \mathcal{P}(\mathbb{R}^d)$. In analogy to Section 2.4.1, a natural approach is to choose functions $\overline{\gamma}_{\alpha,n}$ and $\overline{\eta}_{\alpha,n}$ such that

$$\mathbb{P}\left(F_{\theta,n}^{-1}\big(\overline{\gamma}_{\alpha,n}(u)\big) \leq F_\theta^{-1}(u) \leq F_{\theta,n}^{-1}\big(\overline{\eta}_{\alpha,n}(u)\big),\ \forall u \in (0,1), \theta \in \mathbb{S}^{d-1}\right) \geq 1 - \frac{\alpha}{2}, \tag{2.26}$$

uniformly in $P \in \mathcal{P}(\mathbb{R}^d)$, where $F_{\theta,n}(x) = (1/n)\sum_{i=1}^n I(X_i^\top\theta \leq x)$ for all $x \in \mathbb{R}$ and $\theta \in \mathbb{S}^{d-1}$, and F_θ^{-1} denotes the quantile function of $P_\theta = \pi_{\theta\#}P$. Such a bound can be obtained, for instance, by applying the VC inequality (Vapnik, 2013) to the empirical process indexed by the set of half-spaces in $\mathbb{R}^d$. An assumption-free confidence interval for $\mathrm{SW}_{r,\delta}(P,Q)$

with finite-sample coverage may then be constructed by following the same lines as in the previous section. Due to the uniformity of equation (2.26) over the unit sphere, however, it can be seen that the length of such an interval is necessarily dimension-dependent. In what follows, we instead show that it is possible to obtain a confidence interval with dimension-independent length by exploiting the fact that the Sliced Wasserstein distance is a mean with respect to μ.

Let $\theta_1, \dots, \theta_N$ be an i.i.d. sample from the distribution μ, for some integer $N \geq 1$, and let $\mu_N = (1/N)\sum_{i=1}^N \delta_{\theta_i}$ denote the corresponding empirical measure. Consider the following Monte Carlo approximation of the Sliced Wasserstein distance between the distributions P and Q,

$$\mathrm{SW}_{r,\delta}^{(N)}(P,Q) = \left(\int_{\mathbb{S}^{d-1}} W_{r,\delta}^r(P_\theta, Q_\theta) d\mu_N(\theta)\right)^{\frac{1}{r}} = \left(\frac{1}{N}\sum_{j=1}^N W_{r,\delta}^r(P_{\theta_j}, Q_{\theta_j})\right)^{\frac{1}{r}}.$$

For any $\theta \in \mathbb{S}^{d-1}$, let $[\ell_{N,nm}(\theta), u_{N,nm}(\theta)]$ be the confidence interval in equation (2.25) for $W_{r,\delta}(P_\theta, Q_\theta)$, at level $1-\alpha/N$. Let

$$L_{N,nm} = \int_{\mathbb{S}^{d-1}} \ell_{N,nm}^r(\theta) d\mu_N(\theta), \quad U_{N,nm} = \int_{\mathbb{S}^{d-1}} u_{N,nm}^r(\theta) d\mu_N(\theta),$$

and set

$$C_{nm}^{(N)} = \left[L_{N,nm}^{\frac{1}{r}}, U_{N,nm}^{\frac{1}{r}}\right]. \tag{2.27}$$

By a Bonferroni correction, we obtain conditional coverage of $\mathrm{SW}_{r,\delta}^{(N)}(P,Q)$, i.e. almost surely,

$$\inf_{P,Q\in\mathcal{P}(\mathbb{R}^d)} \mathbb{P}\Big(\mathrm{SW}_{r,\delta}^{(N)}(P,Q) \in C_{nm}^{(N)} \;\Big|\; \theta_1, \dots, \theta_N\Big) \geq 1-\alpha.$$

We further obtain finite-sample coverage of the Sliced Wasserstein distance itself by the following small enlargement of $C_{nm}^{(N)}$.

Proposition 7. Let $b, r \geq 1$ and $\delta \in (0, 1/2)$. Let $(M_N)_{N=1}^\infty$ be a nonnegative sequence such that $M_N \to \infty$ as $N \to \infty$. Define

$$\overline{C}_{nm}^{(N)} = \left[\Big(L_{N,nm} - M_N/\sqrt{N}\Big)^{\frac{1}{r}}, \Big(U_{N,nm} + M_N/\sqrt{N}\Big)^{\frac{1}{r}}\right]. \tag{2.28}$$

Then, there is a constant $c > 0$ depending only on r such that

$$\inf_{P,Q\in\mathcal{K}_{2r}(b)} \mathbb{P}\Big(\mathrm{SW}_{r,\delta}(P,Q) \in \overline{C}_{nm}^{(N)}\Big) \geq 1-\alpha-\frac{bc}{M_N^2\delta^r}.$$

Proposition 7 ensures that an enlargement of the interval $C_{nm}^{(N)}$, of size less than $(M_N/\sqrt{N})^{\frac{1}{r}}$, will cover $\mathrm{SW}_{r,\delta}(P,Q)$ at level $1-\alpha-O(M_N^{-2})$, for any fixed sample sizes n and m. Notice that N is chosen by the practitioner, so that this enlargement can be made to be of lower order than the length of $C_{nm}^{(N)}$. We shall therefore focus our analysis and numerical studies on the interval $C_{nm}^{(N)}$ rather than $\overline{C}_{nm}^{(N)}$.

Although the coverage of the above intervals requires no assumptions on P and Q, apart from the mild moment condition $P, Q \in \mathcal{K}_{2r}(b)$, we now show that their length achieves the minimax rates established in Theorem 5, and is adaptive to the magnitude of $\mathrm{SJ}_{r,\delta}(P), \mathrm{SJ}_{r,\delta}(Q)$.

2.4.2.1 Bounds on the Confidence Interval Length

In this section, we state a general upper bound (Theorem 7) on the length of $C_{nm}^{(N)}$, depending on $\gamma_{\alpha,n}, \eta_{\alpha,n}$. We subsequently specialize this result through Corollaries 1, 2 to illustrate the different rates of convergence which can be obtained under various choices of these functions, and under various conditions on the underlying distributions.

In what follows, we assume $\gamma_{\alpha,n}$ and $\eta_{\alpha,n}$ are both differentiable, invertible with differentiable inverses over $(0,1)$, and are respectively increasing and decreasing as functions of α. Given $\epsilon \in (0,1)$, for notational convenience we write $\varepsilon := (\epsilon \wedge \alpha)/N$ and $a = \alpha/N$. In the sequel, we also omit explicitly indexing various quantities by the number of Monte Carlo samples N. Our upper bounds on the length of $C_{nm}^{(N)}$ will depend on the function

$$\widetilde{\kappa}_{\varepsilon,n}(u) = \max\Big\{|f^{-1}(u) - g^{-1}(u)| : f, g \in \{\gamma_{a,n}, \gamma_{\varepsilon,n}, \eta_{a,n}, \eta_{\varepsilon,n}\}\Big\}, \quad u \in (0,1),$$

as measured by the following two sequences,

$$\kappa_{\varepsilon,n} = \sup_{\frac{\delta}{2} \leq u \leq 1-\frac{\delta}{2}} \widetilde{\kappa}_{\varepsilon,n}(u), \quad V_{\varepsilon,n}(P) = \frac{1}{1-2\delta} \int_{\mathbb{S}^{d-1}} \int_{\delta/2}^{1-\delta/2} \left[\frac{\widetilde{\kappa}_{\varepsilon,n}(u)}{p_\theta(F_\theta^{-1}(u))}\right]^r du d\mu(\theta).$$

Here, recall p_θ denotes the density of the absolutely continuous component of P_θ. Additional technical assumptions **B1**-**B3** regarding $\gamma_{\alpha,n}, \eta_{\alpha,n}, \kappa_{\varepsilon,n}$, appear in Section 2.E. For appropriate choices of δ and ϵ, these assumptions are satisfied by the choices of $\gamma_{\alpha,n}$ and $\eta_{\alpha,m}$ described in Examples 1 and 2, for which the corresponding values of $\kappa_{\varepsilon,n}$ and $V_{\varepsilon,n}(P)$ are derived in the following simple Lemma.

Lemma 1. *Let $\varepsilon \in (0,1)$.*

1. *If $\gamma_{\varepsilon,n}$ and $\eta_{\varepsilon,m}$ are chosen as in equation* (2.22)*, then there exist constants $c_1, c_2 > 0$ depending only on r such that*

$$\kappa_{\varepsilon,n} \leq c_1 \sqrt{\frac{\log(4/\varepsilon)}{n}}, \quad \textit{and} \quad V_{\varepsilon,n}(P) \leq c_2 \left(\frac{\kappa_{\varepsilon,n}}{\sqrt{\delta}}\right)^r \mathrm{SJ}_{r,\frac{\delta}{2}}(P).$$

2. *If $\gamma_{\varepsilon,n}$ and $\eta_{\varepsilon,m}$ are chosen as in equation* (2.24)*, then there exist constants $k_1, k_2 > 0$ depending only on r such that*

$$\kappa_{\varepsilon,n} \leq k_1 \nu_{\varepsilon,n}, \quad \textit{and} \quad V_{\varepsilon,n}(P) \leq k_2 \nu_{\varepsilon,n}^r \mathrm{SJ}_{r,\frac{\delta}{2}}(P).$$

The proof is a straightforward consequence of Examples 1, 2, together with the derivations in Appendices 2.B and 2.G.5, and is therefore omitted. We now define the functional

$$U_{\varepsilon,n}(P) = \frac{1}{1-2\delta}\int_{\mathbb{S}^{d-1}} \left(\sup_{\substack{\delta \leq u \leq 1-\delta \\ |h| \leq \kappa_{\varepsilon,n}}} \left| F_\theta^{-1}(u+h) - F_\theta^{-1}(u) \right|^{r-1} \right) d\mu(\theta).$$

$U_{\varepsilon,n}(P)$ is an upper bound on the magnitude of the largest jump discontinuity of the quantile function F_θ^{-1}, averaged over directions $\theta \in \mathbb{S}^{d-1}$. When $\mathrm{SJ}_{r,\delta}(P) < \infty$, the quantile function F_θ^{-1} is absolutely continuous for almost all $\theta \in \mathbb{S}^{d-1}$ (see Lemma 2 in Section 2.A), implying that $U_{\varepsilon,n}(P)$ decays to zero as $n \to \infty$. The lengths of our confidence intervals will now depend on the quantities

$$\psi_{\varepsilon,nm}=\begin{cases} \big(\mathrm{SW}_{\infty,\delta}^{(r-1)}(P,Q)+U_{\varepsilon,n}(P)+U_{\varepsilon,m}(Q)\big)\frac{\sqrt{b}\kappa_{\varepsilon,n}}{\sqrt{\delta}}, & \mathrm{SJ}_{r,\frac{\delta}{2}}(P)\vee\mathrm{SJ}_{r,\frac{\delta}{2}}(Q)=\infty \\ \big(\mathrm{SW}_{r,\delta}^{r}(P,Q)+V_{\varepsilon,n}(P)+V_{\varepsilon,m}(Q)\big)^{\frac{r-1}{r}}\left[V_{\varepsilon,n}(P)\right]^{\frac{1}{r}}, & \text{otherwise,} \end{cases}$$

and,

$$\varphi_{\varepsilon,nm}=\begin{cases} \big(\mathrm{SW}_{\infty,\delta}^{(r-1)}(P,Q)+U_{\varepsilon,n}(P)+U_{\varepsilon,m}(Q)\big)\frac{\sqrt{b}\kappa_{\varepsilon,m}}{\sqrt{\delta}}, & \mathrm{SJ}_{r,\frac{\delta}{2}}(P)\vee\mathrm{SJ}_{r,\frac{\delta}{2}}(Q)=\infty \\ \big(\mathrm{SW}_{r,\delta}^{r}(P,Q)+V_{\varepsilon,n}(P)+V_{\varepsilon,m}(Q)\big)^{\frac{r-1}{r}}\left[V_{\varepsilon,m}(Q)\right]^{\frac{1}{r}}, & \text{otherwise.} \end{cases}$$

With this notation in place, we arrive at the following upper bound on the length of $C_{nm}^{(N)}$. Recall that λ denotes the Lebesgue measure on $\mathbb{R}$. To simplify our statement, we shall only consider the case where δ is bounded away from $1/2$.

Theorem 7. *Let $r, b \geq 1$ and $\alpha, \epsilon \in (0,1)$. Let $P, Q \in \overline{\mathcal{K}}_2(b)$, and define $\delta \in (0, \delta_0)$ for some $\delta_0 \in (0, 1/2)$. Recall that $\varepsilon = (\epsilon \wedge \alpha)/N$, and assume $\kappa_{\varepsilon,n\wedge m} \leq \frac{\delta}{2} \wedge (1-2\delta)$. Assume further that conditions **B1-B3** hold for some constants $K_1, K_2 > 0$. Then, there exists $c > 0$ depending only on K_1, K_2, δ_0, r such that with probability at least $1-\epsilon$,*

$$\lambda(C_{nm}^{(N)}) \leq \Big\{ \mathrm{SW}_{r,\delta}^r(P,Q) + c\big(\psi_{\varepsilon,nm} + \varphi_{\varepsilon,nm} + \varkappa_N\big)\Big\}^{1/r} - \mathrm{SW}_{r,\delta}(P,Q).$$

Here, $\varkappa_N$ denotes a random variable depending only on μ_N, such that $\mathbb{E}|\varkappa_N| \leq c_1 N^{-1/2r} I(d \geq 2)$, for a constant $c_1 > 0$ depending on b, r, δ and K_1–K_2.

The proof of Theorem 7 appears in Section 2.E. As we shall see, the presence of $\mathrm{SW}_{\infty,\delta}^{(r-1)}(P,Q)$ or $\mathrm{SW}_{r,\delta}(P,Q)$ in the definition of $\varphi_{\varepsilon,nm}$ and $\psi_{\varepsilon,nm}$ implies distinct rates of decay for the confidence interval length, depending on whether P, Q approach each other under the Sliced Wasserstein distance. The fact that $\mathrm{SW}_{\infty,\delta}$ is a stronger metric than $\mathrm{SW}_{r,\delta}$, and the presence of the functional $U_{\varepsilon,n}$, will imply a second dichotomy in the rate of decay of the confidence interval length, based on whether or not $\mathrm{SJ}_{r,\delta/2}(P) \vee \mathrm{SJ}_{r,\delta/2}(Q) < \infty$.

The following result specializes Theorem 7 to Examples 1 and 2.

Corollary 1. *Let $r, b \geq 1$ and $\alpha, \epsilon \in (0,1)$. Let $P, Q \in \overline{\mathcal{K}}_2(b)$, and define $\delta \in (0, \delta_0)$ for some $\delta_0 \in (0, 1/2)$.*

(i) Suppose $\gamma_{\alpha,n}, \eta_{\alpha,n}$ are chosen as in Example 1. Then, there is a constant $c_\alpha > 0$ such that whenever $\delta \wedge (1-2\delta) \geq \sqrt{c_\alpha \log(N/\epsilon)/(n \wedge m)}$, we have with probability at least $1-\epsilon$,

$$\lambda(C_{nm}^{(N)}) \underset{\alpha,b,\delta_0,r}{\lesssim} \varkappa_N^{\frac{1}{r}} + \begin{cases} \delta^{-\frac{1}{2}} \log(N/\epsilon)^{\frac{1}{2r}} \left(n^{-\frac{1}{2r}} + m^{-\frac{1}{2r}}\right), & \mathrm{SJ}_{r,\frac{\delta}{2}}(P) \vee \mathrm{SJ}_{r,\frac{\delta}{2}}(Q) = \infty \\ \delta^{-\frac{1}{2}} \log(N/\epsilon)^{\frac{1}{2}} \Big(\frac{\mathrm{SJ}_{r,\delta/2}^{\frac{1}{r}}(P)}{\sqrt{n}} + \frac{\mathrm{SJ}_{r,\delta/2}^{\frac{1}{r}}(Q)}{\sqrt{m}}\Big), & \text{otherwise.} \end{cases}$$

(ii) Suppose $\gamma_{\alpha,n}, \eta_{\alpha,n}$ are chosen as in Example 2. Let $\beta_{\epsilon,nm} = \log(n \wedge m) + \log(N/\epsilon)$. Then, there is a constant $k_\alpha > 0$ such that whenever $\delta \wedge (1-2\delta) \geq \sqrt{k_\alpha \beta_{\epsilon,nm}/(n \wedge m)}$, we have with probability at least $1-\epsilon$,

$$\lambda(C_{nm}^{(N)}) \underset{\alpha,b,\delta_0,r}{\lesssim} \varkappa_N^{\frac{1}{r}} + \begin{cases} \delta^{-\frac{1}{2}} \beta_{\epsilon,nm}^{\frac{1}{2r}} \left(n^{-\frac{1}{2r}} + m^{-\frac{1}{2r}}\right), & \mathrm{SJ}_{r,\frac{\delta}{2}}(P) \vee \mathrm{SJ}_{r,\frac{\delta}{2}}(Q) = \infty \\ \beta_{\epsilon,nm}^{\frac{1}{2}} \left(\frac{\mathrm{SJ}_{r,\delta/2}^{\frac{1}{r}}(P)}{\sqrt{n}} + \frac{\mathrm{SJ}_{r,\delta/2}^{\frac{1}{r}}(Q)}{\sqrt{m}}\right), & \text{otherwise.} \end{cases}$$

Whenever the trimming sequence is chosen as $\delta \asymp (n \wedge m)^{-a}$ for some $a \in (0, 1/2)$, notice that one may allow ϵ to vanish at an exponentially fast rate with respect to $n \wedge m$, in both cases of Corollary 1. The high-probability bounds in this result may then be turned into bounds on the expected confidence interval lengths, similarly as in Proposition 4, though we avoid doing so here for brevity.

Corollary 1(i) shows that the length of the DKW-based interval achieves the minimax lower bound of Theorem 5(i), up to a polylogarithmic factor in N and the approximation error $\varkappa_N$. It does not, however, achieve the optimal dependence on δ. This is a consequence of the DKW inequality not adapting to the variance of the distributions therein. Corollary 1(ii) instead shows that the relative VC-based interval has length depending on δ solely through the magnitude of the $\mathrm{SJ}_{r,\delta/2}$ functional, at the expense of a polylogarithmic term in n, m. In both cases, the confidence interval length scales polynomially with δ^{-1} when $\mathrm{SJ}_{r,\delta/2}(P) \vee \mathrm{SJ}_{r,\delta/2}(Q) = \infty$, suggesting that in this case, the practitioner should not let δ vanish with $n \wedge m$ at a rate faster than logarithmic, to guarantee consistent inference.

When the distributions P and Q are assumed to be bounded away from each other in $\mathrm{SW}_{r,\delta}$, Theorem 5(ii) suggests that the nonparametric rate $n^{-\frac{1}{2r}} + m^{-\frac{1}{2r}}$ in Corollary 1 is improvable. This is indeed the case, as shown below.

Corollary 2. *Suppose $P, Q \in \overline{\mathcal{K}}_2(b)$ satisfy $\mathrm{SW}_{r,\delta}(P,Q) \geq \Gamma$ for some constant $\Gamma > 0$. Then, under the assumptions of Theorem 7, we have with probability at least $1-\epsilon$,*

$$\lambda(C_{nm}^{(N)}) \underset{\Gamma,\delta_0,b,r}{\lesssim} \varkappa_N + \begin{cases} \delta^{-\frac{r}{2}} \left(\kappa_{\varepsilon,n} + \kappa_{\varepsilon,m}\right), & \mathrm{SJ}_{r,\delta/2}(P) \vee \mathrm{SJ}_{r,\delta/2}(Q) = \infty \\ V_{\varepsilon,n}^{1/r}(P) + V_{\varepsilon,m}^{1/r}(Q), & \text{otherwise.} \end{cases}$$

For example, when $\gamma_{\epsilon,n}, \eta_{\epsilon,n}$ are based on the DKW inequality (Example 1), Corollary 2 implies that the length of $C_{nm}^{(N)}$ achieves the parametric rate $n^{-\frac{1}{2}} + m^{-\frac{1}{2}}$ with high probability (ignoring factors depending only on N and δ), under the mere condition that P and Q are bounded away from each other. Theorem 5(ii) implies that this rate is minimax optimal. As before, adaptivity to the magnitudes of $\mathrm{SJ}_{r,\delta/2}(P), \mathrm{SJ}_{r,\delta/2}(Q)$, without further dependence on δ, is available using the relative VC-based interval in Example 2.

2.5 Asymptotic Confidence Intervals

We now discuss several existing asymptotic confidence intervals for the one-dimensional Wasserstein distance, their extensions to the Sliced Wasserstein distance, and we compare them to our finite-sample confidence intervals in Section 2.4.

In the context of goodness-of-fit testing, Munk and Czado (1998) prove central limit theorems of the form $\sqrt{\frac{nm}{n+m}}\left\{W_{2,\delta}^2(P_n, Q_m) - W_{2,\delta}^2(P,Q)\right\} \rightsquigarrow N(0,\sigma^2)$, where P and Q are one-dimensional distributions, and $\sigma > 0$. They also construct a consistent estimator of σ. These results assume that $P \neq Q$, and that each of P and Q satisfy the following condition,

(C) F is twice continuously differentiable, with density p, which is strictly positive over the real line. Moreover,
$$\sup_{x\in\mathbb{R}} F(x)(1-F(x))\left|\frac{p'(x)}{p^2(x)}\right| < \infty.$$

Assumption (C) originates from strong approximation theorems for the empirical quantile process (Csorgo and Revesz, 1978), and entails that P and Q have differentiable densities, whose supports are intervals. Under the weaker assumption that P and Q merely admit continuous and positive densities on the real line, and still retaining the assumption that $P \neq Q$, Freitag, Munk, and Vogt (2003); Freitag and Munk (2005); Freitag, Czado, and Munk (2007) prove the consistency of the bootstrap in estimating the distribution of $W_{2,\delta}^2(P_n, Q_m)$ in the one-dimensional case.

The Wasserstein distance is well-defined between any pairs of (possibly mutually singular) distributions with sufficient moments, unlike other classical metrics between probability distributions such as the Hellinger and L^r metrics. Indeed, this feature of the Wasserstein distance is a primary motivation for its use in statistical applications. Smoothness assumptions such as (C) are therefore prohibitive in inferential problems for the Wasserstein distance, and motivated our development of assumption-light confidence intervals in the previous section. Nevertheless, when a smoothness assumption such as (C) happens to hold, asymptotic confidence intervals based on limit laws such as those of Munk and Czado (1998) above, or those based on the bootstrap, may have shorter length than those described in Section 2.4.

We show in Section 2.5.1 that under some regularity conditions, the bootstrap is valid in estimating the distribution of $\mathrm{SW}_{r,\delta}(P_n, Q_m)$ for all $r > 1$, thereby generalizing the results of Freitag, Munk, and Vogt (2003); Freitag and Munk (2005); Freitag, Czado, and Munk (2007) from the case $d = 1$ and $r = 2$. We then illustrate in Section 2.5.2, how the strengths of bootstrap can be combined with those of the finite-sample confidence intervals of Section 2.4.

2.5.1 Bootstrapping the Sliced Wasserstein Distance

Let $P, Q \in \mathcal{P}(\mathbb{R}^d)$, and let $X_1, \dots, X_n \sim P$, $Y_1, \dots, Y_m \sim Q$ be i.i.d. samples which are independent of each other. Furthermore, let P_n and Q_m denote their corresponding empirical measures, and let P_n^* and Q_m^* denote their bootstrap counterparts (that is, P_n^* is the sampling distribution of a sample of size n drawn from P_n). Lemma 13 in Section 2.F establishes the Hadamard differentiability of the Sliced Wasserstein distance at pairs of distributions (P, Q) satisfying certain regularity conditions. Limit laws for the empirical Sliced Wasserstein distance, together with consistency of the bootstrap, then follow from the functional delta method (van der Vaart and Wellner, 1996), as outlined in Theorem 8 below. We first introduce some notation. Denote by BL_1 the set of 1-Lipschitz functions $f : \mathbb{R} \to \mathbb{R}$, such that $\|f\|_\infty \leq 1$. We also write for all $u \in [\delta, 1-\delta]$ and all $\theta \in \mathbb{S}^{d-1}$,

$$w(u, \theta) = \frac{r}{1-2\delta} \mathrm{sgn}\left(F_\theta^{-1}(u) - G_\theta^{-1}(u)\right) \left|F_\theta^{-1}(u) - G_\theta^{-1}(u)\right|^{r-1},$$

as well as,

$$\sigma_P^2 = \int_0^{1-\delta} \left(\int_{\mathbb{S}^{d-1}} \int_{F_\theta^{-1}(\delta \vee t)}^{F_\theta^{-1}(1-\delta)} w(F_\theta(x), \theta) dx d\mu(\theta) \right)^2 dt \\ - \left(\int_0^{1-\delta} \int_{\mathbb{S}^{d-1}} \int_{F_\theta^{-1}(\delta \vee t)}^{F_\theta^{-1}(1-\delta)} w(F_\theta(x), \theta) dx d\mu(\theta) dt \right)^2,$$

and,

$$\sigma_Q^2 = \int_0^{1-\delta} \left(\int_{\mathbb{S}^{d-1}} \int_{G_\theta^{-1}(\delta \vee t)}^{G_\theta^{-1}(1-\delta)} w(G_\theta(x), \theta) dx d\mu(\theta) \right)^2 dt \\ - \left(\int_0^{1-\delta} \int_{\mathbb{S}^{d-1}} \int_{G_\theta^{-1}(\delta \vee t)}^{G_\theta^{-1}(1-\delta)} w(G_\theta(x), \theta) dx d\mu(\theta) dt \right)^2,$$

Finally, for any absolutely continuous distribution $P \in \mathcal{P}(\mathbb{R})$ with density p, we shall make use of the following trimmed version of the J_∞ functional introduced by Bobkov and Ledoux (2019),

$$J_{\infty,\delta}(P) = \operatorname*{esssup}_{\delta \leq u \leq 1-\delta} \frac{1}{p(F^{-1}(u))}.$$

With this notation in place, the main result of this section is stated as follows.

Theorem 8. *Let $\delta \in [0, 1/2)$ and $r > 1$. Let $P, Q \in \overline{\mathcal{K}}_2$ be distributions such that for all $\theta \in \mathbb{S}^{d-1}$, P_θ, Q_θ are absolutely continuous with respect to the Lebesgue measure, with respective families of densities $\{p_\theta\}_{\theta \in \mathbb{S}^{d-1}}$, $\{q_\theta\}_{\theta \in \mathbb{S}^{d-1}}$ which are uniformly integrable over $\mathbb{R}$. Assume that,*

$$\sup_{\theta \in \mathbb{S}^{d-1}} J_{\infty,\delta/2}(P_\theta) \vee J_{\infty,\delta/2}(Q_\theta) < \infty. \tag{2.29}$$

Then, the following statements hold as $n, m \to \infty$ such that $\frac{n}{n+m} \to a \in (0, 1)$.

(i) (Central Limit Theorem) We have,

$$\sqrt{\frac{nm}{n+m}}\Big(\mathrm{SW}_{r,\delta}^r(P_n, Q_m) - \mathrm{SW}_{r,\delta}^r(P, Q)\Big) \rightsquigarrow N\big(0, (1-a)\sigma_P^2 + a\sigma_Q^2\big).$$

(ii) (Bootstrap Consistency) For $\mathbf{X} = (X_1, \dots, X_n), \mathbf{Y} = (Y_1, \dots, Y_m)$, we have

$$\sup_{h \in \mathrm{BL}_1} \Bigg| \mathbb{E}\left[h\left(\sqrt{\frac{nm}{n+m}} \left\{ \mathrm{SW}_{r,\delta}^r(P_n^*, Q_m^*) - \mathrm{SW}_{r,\delta}^r(P_n, Q_m) \right\} \right) \middle| \mathbf{X}, \mathbf{Y} \right] - \mathbb{E}\left[h\left(\sqrt{\frac{nm}{n+m}} \left\{ \mathrm{SW}_{r,\delta}^r(P_n, Q_m) - \mathrm{SW}_{r,\delta}^r(P, Q) \right\} \right) \right] \Bigg| \to 0,$$

in outer probability.

Theorem 8(i) provides a central limit theorem for the empirical trimmed Sliced Wasserstein distance, centered at its population counterpart. The primary assumptions required for this result are (a) the existence and uniform integrability of the densities of P_θ and Q_θ along directions $\theta \in \mathbb{S}^{d-1}$, and (b) a uniform lower bound on these densities, over the compact sets $[F_\theta^{-1}(\delta/2), F_\theta^{-1}(1-\delta/2)]$, as measured by the $J_{\infty,\delta/2}$ functional. Note that assumption (a) holds if P, Q admit upper bounded densities with respect to the Lebesgue measure over $\mathbb{R}^d$, but is strictly weaker; indeed, it can be satisfied by non-atomic measures which are singular with respect to the Lebesgue measure on $\mathbb{R}^d$. Furthermore, we note that the assumption of uniform integrability is vacuous in the special case $d = 1$. Assumption (b) requires the bulk of the supports of P_θ and Q_θ to be connected. Such a condition is necessary for the limit in Theorem 8(i) to be a mean-zero Gaussian distribution, as can be anticipated, for instance, from the lack of Hadamard differentiability of the Wasserstein distance over finite spaces (Sommerfeld and Munk, 2018). Nevertheless, our condition is stronger, since we have assumed $J_{\infty,\delta/2}(P_\theta)$ and $J_{\infty,\delta/2}(Q_\theta)$ are uniformly bounded in θ. Inspired by our results on estimation and finite sample inference for the Sliced Wasserstein distance, it is natural to ask whether this condition can be replaced by, say, $\mathrm{SJ}_{r,\delta}(P), \mathrm{SJ}_{r,\delta}(Q) < \infty$. Such a condition would allow the densities p_θ and q_θ to approach zero at a sufficiently slow rate, which is currently precluded by our theorem. We leave this question open for future work.

In the special case $d = 1$ and $r = 2$, the limiting variance obtained in Theorem 8(i) is equal to the one obtained by Munk and Czado (1998), up to renormalizing their definition of the trimmed Wasserstein distance, though our assumptions are significantly weaker since we do not require the aforementioned condition (C). Nevertheless, their result allows the trimming constant δ to vanish, while we require δ to be held fixed and, in fact, positive, when P and Q have unbounded support. In this regard, our result is closer to those of Freitag, Munk, and Vogt (2003); Freitag and Munk (2005); Freitag, Czado, and Munk (2007), who prove, in particular, the Hadamard differentiability of the functional $W_{2,\delta}^2$ for a nonvanishing trimming constant δ. Their results require P and Q to admit positive and continuously differentiable densities over the real line, which is a strictly stronger assumption than those of Theorem 8. In particular, we require no smoothness conditions on the various densities.

We next compare Theorem 8(i) to existing central limit theorems for untrimmed Wasserstein distances. Let $d = 1$, and assume P and Q are compactly-supported, so that one may take $\delta = 0$ in Theorem 8(i). In this case, the limiting variance may be reformulated in terms of the expressions

$$\sigma_P^2 = \mathrm{Var}[\phi_0(X)], \quad \sigma_Q^2 = \mathrm{Var}[\psi_0(Y)],$$

where $X \sim P \in \mathcal{P}(\mathbb{R})$, $Y \sim Q \in \mathcal{P}(\mathbb{R})$, and for all $x, y \in \mathbb{R}$,

$$\phi_0(x) = \int_{-\infty}^{x} w(F(t))dt, \quad \psi_0(y) = \int_{-\infty}^{y} w(G(t))dt.$$

Here, we abbreviate $w(\cdot) \equiv w(\cdot, \theta)$ in the one-dimensional case. It can be deduced from Gangbo and McCann (1996) that (ϕ_0, ψ_0) forms an optimal pair of Kantorovich potentials in the dual $|\cdot|^r$-optimal transport problem (Villani, 2003) from P to Q. In particular, for $r = 2$, the limiting variance in Theorem 8(i) reduces to the one obtained by del Barrio and Loubes (2019), who derive central limit theorems for $W_2^2(P_n, Q_m)$, in general dimension $d \geq 1$. Their results are not centered at $W_2^2(P, Q)$ due to the large bias of empirical Wasserstein distances in general dimension, however, it was shown by del Barrio, Gordaliza, and Loubes (2019) that when $d = 1$, these limit theorems may be centered at the population Wasserstein distance under assumptions akin to condition (C). We also refer to Berthet, Fort, and Klein (2020) and the recent work of Hundrieser et al. (2022) for distinct assumptions under which such a result can be obtained.

Theorem 8(ii) proves the consistency of the bootstrap in estimating the distribution of $\mathrm{SW}_{r,\delta}^r(P_n, Q_m)$. Letting F_{nm}^* denote the CDF of $\mathrm{SW}_{r,\delta}^r(P_n^*, Q_m^*) - \mathrm{SW}_{r,\delta}^r(P_n, Q_m)$, it follows that an asymptotic $(1 - \alpha)$-confidence interval for $\mathrm{SW}_{r,\delta}(P, Q)$ is given by

$$C_{nm}^* = \left[\left(\mathrm{SW}_{r,\delta}^r(P_n, Q_m) - F_{nm}^*(1 - \alpha/2)\right)^{\frac{1}{r}}, \left(\mathrm{SW}_{r,\delta}^r(P_n, Q_m) + F_{nm}^*(\alpha/2)\right)^{\frac{1}{r}}\right].$$

The CDF F_{nm}^* is typically estimated via Monte Carlo simulation (Efron and Tibshirani, 1994). The assumptions for the validity of C_{nm}^* are those of Theorem 8, and in addition, the condition that $\mathrm{SW}_{r,\delta}(P, Q) > 0$, which is necessary and sufficient for the limiting variance $a\sigma_P^2 + (1 - a)\sigma_Q^2$ in Theorem 8 to be positive. Failure of the bootstrap at the null $\mathrm{SW}_{r,\delta}(P, Q) = 0$ is due to the Sliced Wasserstein distance being a functional with first-order degeneracy (Munk and Czado, 1998), for which corrections such as those of Chen and Fang (2019), or the m-out-of-n bootstrap (Sommerfeld and Munk, 2018), yield consistent procedures, but are practically less attractive as they introduce further tuning parameters.

We illustrate in the sequel how our finite sample confidence intervals can be combined with the bootstrap to relax this assumption.

2.5.2 A Hybrid Bootstrap Approach

Let C_{nm}^* denote the preceding bootstrap confidence interval at level $1 - \alpha/2$, and let $C_{nm}^\dagger$ denote the assumption-light confidence interval for $\mathrm{SW}_{r,\delta}(P, Q)$ in equation (2.28) at level

$1-\alpha/2$. Assume that the number of Monte Carlo replications N therein is taken to diverge as $n,m\to\infty$. We define the $(1-\alpha)$-hybrid confidence interval as:

$$C_{nm}=\begin{cases}C_{nm}^\dagger, & \text{if } 0\in C_{nm}^\dagger,\\ C_{nm}^*, & \text{otherwise.}\end{cases} \tag{2.30}$$

Roughly, we use the bootstrap interval if we are reasonably certain that P and Q are bounded away from each other in Sliced Wasserstein distance, and fall back on the finite-sample interval otherwise. The following simple result characterizes the asymptotic coverage and length of the hybrid interval. In order to simplify our discussion, we write $C_{nm}=[a_{nm}^{1/r},b_{nm}^{1/r}]$ and we focus on bounding the length of the confidence interval $[a_{nm},b_{nm}]$ for the r-th power of the r-Sliced Wasserstein distance. We also assume that the finite-sample interval $C_{nm}^\dagger$ is defined in terms of the DKW confidence band in Example 1.

Proposition 8. Let $a,\alpha\in(0,1)$, $\delta\in(0,1/2)$, and assume the same conditions as Theorem 8. Then, the following holds assuming $\frac{n}{n+m}\to a$ when $n,m\to\ \infty$.

(i) (Coverage) We have,

$$\liminf_{n,m\to\infty}\mathbb{P}\big(\mathrm{SW}_{r,\delta}(P,Q)\in C_{nm}\big)\geq 1-\alpha. \tag{2.31}$$

(ii) (Length) Let $N\asymp n^{r^2}$, and choose $M_N\asymp\log N$ in the definition of $C_{nm}^\dagger$. Then, we have with probability at least $1-\alpha$,

$$(b_{nm}-a_{nm})=O\left(\left(\frac{\log n}{n}\right)^{\frac{r}{2}}+\frac{\mathrm{SW}_{r,\delta}(P,Q)}{\sqrt{n}}\right).$$

Proposition 8 establishes the asymptotic coverage of C_{nm} under the same conditions as Theorem 8. In particular, it removes the assumption $\mathrm{SW}_{r,\delta}(P,Q)>0$, which is needed for the asymptotic coverage of the bootstrap interval C_{nm}^*. We note that many other existing corrections of the bootstrap for functionals with first-order degeneracy, such as the m-out-of-n bootstrap or the procedures outlined in Section 2.1 of Verdinelli and Wasserman (2024), involve expanding the asymptotic length of the interval, leading to a loss of efficiency. In contrast, Proposition 8 shows that with high (albeit, fixed) probability, the hybrid interval achieves the rate-optimal asymptotic length both at the null $\mathrm{SW}_{r,\delta}(P,Q)=0$ and away from the null, up to a polylogarithmic factor in N (which can be removed when $d=1$). We emphasize that this adapativity is obtained without tuning parameters, apart from the sequences M_N,N whose precise choice does not greatly alter the properties of C_{nm}. Once again, the choice of these sequences is vacuous in the special case $d=1$.

Though this methodology inherits benefits from both the bootstrap and finite-sample confidence intervals, it is not assumption-free. In principle, it is possible to extend this procedure by empirically testing whether the conditions of Theorem 8 are met, and to use the outcome of such a test in the conditions of equation (2.30). While doing so may allow for certain assumptions to be relaxed, it could become impractical: for instance, we do not know of a test for the finiteness of the $J_{\infty,\delta/2}$ functional which is free of tuning parameters.

2.6 Conclusion and Discussion

Our aim in this chapter has been to develop assumption-light finite-sample confidence intervals for the Sliced Wasserstein distance. After deriving minimax rates for estimating the Sliced Wasserstein distance, which are of independent interest, we bounded the length of our confidence intervals, showing that they achieve near minimax optimal length. Their length is also shown to be adaptive to whether or not the underlying distributions are near the classical null, as well as to their regularity, as measured by the magnitude of the functional $\mathrm{SJ}_{r,\delta}$. These findings contrast asymptotic methods such as the bootstrap, whose validity we show is subject to certain prohibitive assumptions on the underlying distributions, and whose asymptotic length does not enjoy the same adaptivity as that of our finite-sample intervals.

This chapter leaves open the problem of statistical inference for Wasserstein distances in dimension greater than one, which we will return to in later chapters below. Indeed, our work has hinged upon the representation of the one-dimensional Wasserstein distance as the L^r distance between quantile functions, which is unavailable in general dimension. For the same reason, our work does not shed light on statistical inference for other modifications of the Wasserstein distance based on projections of distributions to low-dimensions greater than one, such as those summarized in Section 2.2.3. We have shown that the Sliced Wasserstein distance can be estimated at dimension-independent rates, and it is of interest to understand how this finding changes for other low-dimensional modifications of the Wasserstein distance.

2.A Preliminary Technical Results

In this section, we collect several preliminary results which will frequently be used in the sequel. We begin with the following straightforward Lemma, which follows from Appendix A of Bobkov and Ledoux (2019).

Lemma 2. *Let $P \in \mathcal{P}(\mathbb{R}^d)$, $r \geq 1$, and $\delta \in (0,1/2)$. Let F_θ^{-1} denote the quantile function of $P_\theta = \pi_{\theta\#}P$ for all $\theta \in \mathbb{S}^{d-1}$. If $\mathrm{SJ}_{r,\delta}(P) < \infty$, then $F_\theta^{-1}|_{[\delta,1-\delta]}$ is absolutely continuous for μ-almost all $\theta \in \mathbb{S}^{d-1}$.*

Furthermore, we describe the following characterization of distributions falling in the collections $\mathcal{K}_{r,\rho}(b)$ and $\overline{\mathcal{K}}_r(b)$.

Lemma 3. *Let $\delta \in (0,1/2)$ and $r, \rho, b \geq 1$. Then, for all distributions $P \in \mathcal{K}_{r,\rho}(b)$,*

$$\int_{\mathbb{S}^{d-1}} \left|F_\theta^{-1}(a)\right|^r d\mu(\theta) \leq b(2/\delta)^{r/\rho}, \quad a \in \{\delta, 1-\delta\}.$$

Furthermore, for all distributions $P \in \overline{\mathcal{K}}_r(b)$, we have

$$\sup_{\theta\in\mathbb{S}^{d-1}} \left|F_\theta^{-1}(a)\right| \leq \left(\frac{2b}{\delta}\right)^{\frac{1}{r}}, \quad a \in \{\delta, 1-\delta\}.$$

Proof of Lemma 3. The claim is a simple consequence of Markov's inequality. Given $\theta \in \mathbb{S}^{d-1}$, it must hold that (a) $F_\theta^{-1}(\delta) < 0$ or (b) $F_\theta^{-1}(1-\delta) > 0$. In the former case, since

$F_\theta(F_\theta^{-1}(\delta)) \geq \delta$, we have

$$\begin{aligned}\delta \leq \mathbb{P}\big\{X^\top\theta \leq F_\theta^{-1}(\delta)\big\} &= \mathbb{P}\big\{-X^\top\theta \geq |F_\theta^{-1}(\delta)|\big\} \\ &\leq \mathbb{P}\big\{|X^\top\theta| \geq |F_\theta^{-1}(\delta)|\big\} \leq \frac{\mathbb{E}_X[|X^\top\theta|^\rho]}{|F_\theta^{-1}(\delta)|^\rho}. \end{aligned} \tag{2.32}$$

while in the latter case, $F_\theta(F_\theta^{-1}(1-\delta)/2^{1/\rho}) \leq 1-\delta$, therefore

$$\delta \leq \mathbb{P}\left\{X^\top\theta \geq \frac{F_\theta^{-1}(1-\delta)}{2^{1/\rho}}\right\} \leq \mathbb{P}\left\{|X^\top\theta| \geq \frac{|F_\theta^{-1}(1-\delta)|}{2^{1/\rho}}\right\} \leq \frac{2\mathbb{E}[|X^\top\theta|^\rho]}{|F_\theta^{-1}(1-\delta)|^\rho} \tag{2.33}$$

We deduce,

$$\max_{a\in\{\delta,1-\delta\}} \left|F_\theta^{-1}(a)\right| \leq \left(\frac{2\mathbb{E}[|X^\top\theta|^\rho]}{\delta}\right)^{\frac{1}{\rho}}. \tag{2.34}$$

Indeed, when both (a) and (b) hold, the above display follows from equations (2.32) and (2.33). When only (a) holds, it is clear that $|F_\theta^{-1}(\delta)| \geq |F_\theta^{-1}(1-\delta)|$, thus the above display follows from equation (2.33), and similarly when only case (b) holds. Thus, since the above display holds for any $\theta \in \mathbb{S}^{d-1}$, we deduce that for both $a \in \{\delta, 1-\delta\}$,

$$\int_{\mathbb{S}^{d-1}} \left|F_\theta^{-1}(a)\right|^r d\mu(\theta) \leq \left(\frac{2}{\delta}\right)^{\frac{r}{\rho}} \int_{\mathbb{S}^{d-1}} \mathbb{E}[|X^\top\theta|^\rho]^{\frac{r}{\rho}} d\mu(\theta) \leq \left(\frac{2}{\delta}\right)^{\frac{r}{\rho}} b,$$

where we used the assumption $P \in \mathcal{K}_{r,\rho}(b)$.

To prove the second claim, one similarly has the following bound when $P \in \overline{\mathcal{K}}_\rho(b)$, for $a \in \{\delta, 1-\delta\}$,

$$\begin{aligned}\sup_{\theta\in\mathbb{S}^{d-1}} \left|F_\theta^{-1}(a)\right| &\leq \sup_{\theta\in\mathbb{S}^{d-1}} \left(\frac{2\mathbb{E}[|X^\top\theta|^\rho]}{\delta}\right)^{\frac{1}{\rho}} \\ &\leq \left(\frac{2}{\delta}\mathbb{E}\left[\sup_{\theta\in\mathbb{S}^{d-1}} |X^\top\theta|^\rho\right]\right)^{\frac{1}{\rho}} = \left(\frac{2}{\delta}\mathbb{E}\left[\|X\|^\rho\right]\right)^{\frac{1}{\rho}} \leq \left(\frac{2b}{\delta}\right)^{\frac{1}{\rho}}.\end{aligned}$$

□

2.B Proof of Propositions 2 and 4

We shall begin by proving Proposition 2(i) and Proposition 4(ii). As we shall explain, Proposition 2(ii) will then follow as a special case of Proposition 4(ii), and Proposition 4(i) will follow from Proposition 2(i). Our proofs will make use of Examples 1 and 2 which appear in Section 2.4 of the main text, and of their corresponding proofs in Section 2.G. We shall also make use of the following Lemma, which is proven in Section 2.B.1.

Lemma 4. *Let $\delta \in (0, 1/2)$. Then, for any $\bar{r} \in [r, r+1]$, there exists a constant $B_r > 0$ depending only on r such that for all $\theta \in \mathbb{S}^{d-1}$ and all $\delta \geq 2(r+2)/n$,*

$$\max_{a\in\{\delta,1-\delta\}} \mathbb{E}\left[\left|F_{\theta,n}^{-1}(a)\right|^{\bar{r}}\right] \leq B_r \left(1 + \max_{a\in\left\{\frac{\delta}{2},1-\frac{\delta}{2}\right\}} \left|F_{\theta}^{-1}(a)\right|^{\bar{r}} + \left(\frac{\mathbb{E}\left|X^\top\theta\right|^2}{\delta}\right)^{\frac{\bar{r}}{2}}\right).$$

In the above result and throughout this section, recall that $F_{\theta,n}^{-1}$ (resp. $G_{\theta,n}^{-1}$, F_θ^{-1}, G_θ^{-1}) denotes the quantile function of the distribution $P_{\theta,n} = \pi_{\theta\#}P_n$ (resp. $Q_{\theta,m} = \pi_{\theta\#}Q_m$, $P_\theta = \pi_{\theta\#}P$, $Q_\theta = \pi_{\theta\#}Q$), for any $\theta \in \mathbb{S}^{d-1}$. Finally, throughout the remainder of this section, the symbol "$\lesssim$" is used to hide universal constants depending only on r.

Proof of Proposition 2(i). The claim is trivial if $\mathrm{SJ}_{r,\delta/2}(P) = \infty$, thus assume otherwise. We shall begin bounding the expectation of the following quantity

$$Z_n(\theta) = W_{r,\delta}^r(P_{\theta,n}, P_\theta),$$

for any fixed $\theta \in \mathbb{S}^{d-1}$. The result will then follow by integration over $\mathbb{S}^{d-1}$. We begin with the following key high probability bound.

Lemma 5. *Let $y_0 = \sqrt{\delta}/4$ and $C_r > 0$ a constant depending only on r. Then, for all $y \in (0, y_0]$, we have*

$$\operatorname*{esssup}_{\theta\in\mathbb{S}^{d-1}} \mathbb{P}\Big(Z_n(\theta) \geq C_r y^r J_{r,\delta/2}(P_\theta)\Big) \leq \frac{2n+1}{16} e^{-\frac{ny^2}{16}},$$

where the essential supremum is taken with respect to the measure μ.

Lemma 5 is proven in Section 2.B.2. Now, let $T_\theta = C_r y_0^r J_{r,\delta/2}(P_\theta)$, so that

$$\mathbb{E}[Z_n(\theta)] = \mathbb{E}[Z_n(\theta) \cdot I(Z_n(\theta) > T_\theta)] + \mathbb{E}[Z_n(\theta) \cdot I(Z_n(\theta) \leq T_\theta)]. \tag{2.35}$$

To bound the first term, notice first that

$$Z_n(\theta) = \frac{1}{1-2\delta}\int_\delta^{1-\delta} \left|F_{\theta,n}^{-1}(u) - F_\theta^{-1}(u)\right|^r du \lesssim \max_{a\in\{\delta,1-\delta\}} \left|F_{\theta,n}^{-1}(a)\right|^r + \left|F_\theta^{-1}(a)\right|^r.$$

Now, let $s = 1/(1-1/\eta)$, where $\eta = \bar{r}/r$ and $\bar{r} \in (r, r+1]$. Then, by Hölder's inequality, we have uniformly in $\theta \in \mathbb{S}^{d-1}$,

$$\begin{aligned}
&\mathbb{E}\big[Z_n(\theta) \cdot I(Z_n(\theta) > T_\theta)\big] \\
&\quad \leq \|Z_n(\theta)\|_{L^\eta(\mathbb{P})} \|I(Z_n(\theta) > T_\theta)\|_{L^s(\mathbb{P})} \\
&\quad \lesssim \max_{a\in\{\delta,1-\delta\}} \left(\left\|F_{\theta,n}^{-1}(a)\right\|_{L^{\bar{r}}(\mathbb{P})}^{\bar{r}} + \left|F_\theta^{-1}(a)\right|^{\bar{r}}\right)^{\frac{1}{\eta}} \|I(Z_n(\theta) > T_\theta)\|_{L^s(\mathbb{P})} \\
&\quad \lesssim \left(1 + \max_{a\in\left\{\frac{\delta}{2},1-\frac{\delta}{2}\right\}} \left|F_\theta^{-1}(a)\right|^{\bar{r}} + \left(\frac{\mathbb{E}|X^\top\theta|^2}{\delta}\right)^{\frac{\bar{r}}{2}}\right)^{\frac{1}{\eta}} \left(\frac{2n+1}{16}\right)^{\frac{1}{s}} e^{-\frac{ny_0^2}{16s}}
\end{aligned} \tag{2.36}$$

$$\leq \left(1 + \max_{a \in \left\{\frac{\delta}{2}, 1-\frac{\delta}{2}\right\}} \left|F_\theta^{-1}(a)\right|^r + \left(\frac{\mathbb{E}|X^\top\theta|^2}{\delta}\right)^{\frac{r}{2}}\right) \left(\frac{2n+1}{16}\right)^{\frac{1}{s}} e^{-\frac{ny_0^2}{16s}},$$

where we invoked Lemmas 3–5 in equation (2.36). Therefore, using the fact that $s \geq 1$, $P \in \mathcal{K}_r(b)$, and invoking Lemma 3, we obtain

$$\int_{\mathbb{S}^{d-1}} \mathbb{E}\big[Z_n(\theta) \cdot I(Z_n(\theta) > T_\theta)\big] d\mu(\theta) \lesssim (b/\delta^{r/2}) n e^{-\frac{ny_0^2}{16s}} \lesssim bne^{-c_0 n\delta}/\delta^{r/2}, \tag{2.37}$$

for a universal constant $c_0 > 0$ depending only on r. We further bound the second term in equation (2.35). Setting $t_\theta = T_\theta \wedge C_r J_{r,\delta/2}(P_\theta) \left(\frac{16}{n} \log\left(\frac{2n+1}{16}\right)\right)^{r/2}$, we arrive at

$$\begin{aligned}
&\mathbb{E}\big[Z_n(\theta) \cdot I(Z_n(\theta) < T_\theta)\big] \\
&\qquad = \int_0^\infty \mathbb{P}\big(Z_n(\theta) \cdot I(Z_n(\theta) < T_\theta) \geq x\big) dx \\
&\qquad \leq \int_0^{T_\theta} \mathbb{P}\big(Z_n(\theta) \geq x\big) dx \\
&\qquad \leq t_\theta + \int_{t_\theta}^{T_\theta} \mathbb{P}\big(Z_n(\theta) \geq x\big) dx \\
&\qquad \leq t_\theta + \frac{2n+1}{16} \int_{t_\theta}^{T_\theta} \exp\left\{-\frac{n}{16}\left(\frac{x}{J_{r,\delta/2}(P_\theta) C_r}\right)^{2/r}\right\} dx \\
&\qquad = t_\theta + \frac{r(2n+1) C_r J_{r,\delta/2}(P_\theta)}{16} \left(\frac{4}{\sqrt{n}}\right)^r \int_{\sqrt{\log\left(\frac{2n+1}{16}\right)}}^\infty e^{-y^2} y^{r-1} dy \\
&\qquad \lesssim t_\theta + \frac{r(2n+1) C_r J_{r,\delta/2}(P_\theta)}{16} \left(\frac{4}{\sqrt{n}}\right)^r \int_{\sqrt{\log\left(\frac{2n+1}{16}\right)}}^\infty e^{-y^2/2} dy,
\end{aligned}$$

where we used the change of variable $y = \frac{\sqrt{n}}{4}\left(\frac{x}{J_{r,\delta}(P_\theta) C_r}\right)^{1/r}$, and where the final inequality holds for all n larger than a universal constant depending only on r. It follows that

$$\mathbb{E}\big[Z_n(\theta) \cdot I(Z_n(\theta) < T_\theta)\big] \lesssim t_\theta + \frac{J_{r,\delta/2}(P_\theta)}{n^{r/2}} \lesssim J_{r,\delta/2}(P_\theta) \left[\frac{1}{n} \log\left(\frac{2n+1}{16}\right)\right]^{r/2}.$$

Putting this fact together with equation (2.37), we have by the Fubini-Tonelli Theorem,

$$\begin{aligned}
\mathbb{E}\big[\mathrm{SW}_{r,\delta}^r(P_n, P)\big] &= \int_{\mathbb{S}^{d-1}} \mathbb{E}[Z_n(\theta)] d\mu(\theta) \\
&\lesssim \mathrm{SJ}_{r,\delta/2}(P) \left(\frac{\log n}{n}\right)^{\frac{r}{2}} + bne^{-c_0 n\delta}/\delta^{r/2}.
\end{aligned}$$

The claim now follows since $\mathbb{E}\big[\mathrm{SW}_{r,\delta}(P_n, P)\big] \leq \Big(\mathbb{E}\big[\mathrm{SW}_{r,\delta}^r(P_n, P)\big]\Big)^{\frac{1}{r}}$. □

Proof of Proposition 4(ii). As we shall show, the assumption $\mathrm{SW}_{r,\delta}(P,Q) \geq \Gamma$ implies that the deviation

$$\Delta_{nm} = \mathrm{SW}_{r,\delta}(P_n, Q_m) - \mathrm{SW}_{r,\delta}(P,Q)$$

is of same order as

$$D_{nm} = \mathrm{SW}^r_{r,\delta}(P_n, Q_m) - \mathrm{SW}^r_{r,\delta}(P,Q).$$

To bound this quantity, define for all $\theta \in \mathbb{S}^{d-1}$,

$$\begin{aligned}
a_\theta &= \min\Big\{F_\theta^{-1}(\delta), F_{\theta,n}^{-1}(\delta), G_\theta^{-1}(\delta), G_{\theta,m}^{-1}(\delta)\Big\},\\
b_\theta &= \max\Big\{F_\theta^{-1}(1-\delta), F_{\theta,n}^{-1}(1-\delta), G_\theta^{-1}(1-\delta), G_{\theta,m}^{-1}(1-\delta)\Big\},\\
M(\theta) &= \max\Big\{\big|F_\theta^{-1}(\delta/2)\big|, \big|G_\theta^{-1}(\delta/2)\big|, \big|F_\theta^{-1}(1-\delta/2)\big|, \big|G_\theta^{-1}(1-\delta/2)\big|, 1\Big\},
\end{aligned}$$

and

$$Z_{nm}(\theta) = (b_\theta - a_\theta)^{r-1}\int_\delta^{1-\delta}\Big[|F_{\theta,n}^{-1}(u) - F_\theta^{-1}(u)| + |G_{\theta,m}^{-1}(u) - G_\theta^{-1}(u)|\Big]du.$$

The bulk of our proof is now contained in the following result.

Lemma 6. *We have,*

$$|D_{nm}| \leq \frac{r}{1-2\delta}\int_{\mathbb{S}^{d-1}} Z_{nm}(\theta)d\mu(\theta).$$

Assume further that $P, Q \in \mathcal{K}_r(b)$. Let $y_0 = \delta/2$. Then, there exists a universal constant $A > 0$ such that for all $y \in (0, y_0]$

$$\sup_{\theta\in\mathbb{S}^{d-1}} \mathbb{P}(|Z_{nm}(\theta)| \geq AM^r(\theta)y) \leq 4\exp(-2(n\wedge m)y^2).$$

Let $A, y_0 > 0$ be as in Lemma 6. Similarly as in the proof of Proposition 2, let $T_\theta = AM^r(\theta)y_0$. Then, we have,

$$\mathbb{E}[Z_{nm}(\theta)] = \mathbb{E}\big[Z_{nm}(\theta)I(Z_{nm}(\theta) \geq T_\theta)\big] + \mathbb{E}\big[Z_{nm}(\theta)I(Z_{nm}(\theta) < T_\theta)\big]. \tag{2.38}$$

Set $s = 1/(1-1/\eta)$, where again $\eta = \bar{r}/r$ and $\bar{r} \in (r, r+1]$, so that by Hölder's inequality, we have uniformly in $\theta \in \mathbb{S}^{d-1}$,

$$\begin{aligned}
&\mathbb{E}\big[Z_{nm}(\theta)\cdot I(Z_{nm}(\theta) \geq T_\theta)\big]\\
&\leq \|Z_{nm}(\theta)\|_{L^\eta(\mathbb{P})}\,\|I(Z_{nm}(\theta) \geq T_\theta)\|_{L^s(\mathbb{P})}\\
&\lesssim \Big(\,\|a_\theta\|^{\bar{r}}_{L^{\bar{r}}(\mathbb{P})} + \|b_\theta\|^{\bar{r}}_{L^{\bar{r}}(\mathbb{P})}\Big)^{\frac{1}{\eta}} \exp(-2(n\wedge m)y_0^2/s)\\
&\lesssim \left(M^{\bar{r}}(\theta) + \left(\frac{\mathbb{E}|X^\top\theta|^2}{\delta}\right)^{\frac{\bar{r}}{2}} + \left(\frac{\mathbb{E}|Y^\top\theta|^2}{\delta}\right)^{\frac{\bar{r}}{2}}\right)^{\frac{1}{\eta}} \exp(-2(n\wedge m)y_0^2/s)
\end{aligned}$$

$$\lesssim \left(M^r(\theta) + \left(\frac{\mathbb{E}|X^\top\theta|^2}{\delta}\right)^{\frac{r}{2}} + \left(\frac{\mathbb{E}|Y^\top\theta|^2}{\delta}\right)^{\frac{r}{2}} \right) \exp(-2(n\wedge m)y_0^2/s), \tag{2.39}$$

where we invoked Lemmas 4 and 6. Now, notice that $\int_{\mathbb{S}^{d-1}} M^r(\theta)d\mu(\theta) \lesssim b/\delta^{r/2}$ by Lemma 3, since $P, Q \in \mathcal{K}_r(b)$. Therefore, integrating both sides of the above display with respect to μ leads to

$$\int_{\mathbb{S}^{d-1}} \mathbb{E}\big[Z_{nm}(\theta)I(Z_{nm}(\theta) \geq T_\theta)\big]d\mu(\theta) \lesssim \exp(-c_1(n\wedge m)\delta^2)b/\delta^{r/2},$$

for a constant $c_1 > 0$. We now bound the second term in equation (2.38). We again use Lemma 6 to obtain

$$\begin{aligned}
&\int_{\mathbb{S}^{d-1}} \mathbb{E}\big[Z_{nm}(\theta)\cdot I(Z_{nm}(\theta) < T_\theta)\big]d\mu(\theta) \\
&\qquad = \int_{\mathbb{S}^{d-1}}\int_0^\infty \mathbb{P}\big(Z_{nm}(\theta)\cdot I(Z_{nm}(\theta) < T_\theta) \geq x\big)dxd\mu(\theta) \\
&\qquad \leq \int_{\mathbb{S}^{d-1}}\int_0^{T_\theta} \mathbb{P}\big(Z_{nm}(\theta) \geq x\big)dxd\mu(\theta) \\
&\qquad \leq 4\int_{\mathbb{S}^{d-1}}\int_0^{T_\theta} \exp\left\{-2(n\wedge m)\left(\frac{x}{AM^r(\theta)}\right)^2\right\}dxd\mu(\theta) \\
&\qquad \lesssim \frac{1}{\sqrt{n\wedge m}}\int_{\mathbb{S}^{d-1}} M^r(\theta)d\mu(\theta) \lesssim \frac{b}{\delta^{r/2}\sqrt{n\wedge m}}.
\end{aligned}$$

Putting this fact together with equation (2.39), and applying the Fubini-Tonelli Theorem and Lemma 6, we arrive at

$$\begin{aligned}
\mathbb{E}|D_{nm}| &\lesssim \frac{1}{1-2\delta}\mathbb{E}\left[\int_{\mathbb{S}^{d-1}} Z_{nm}(\theta)d\mu(\theta)\right] \\
&= \frac{1}{1-2\delta}\int_{\mathbb{S}^{d-1}} \mathbb{E}\big[Z_{nm}(\theta)\big]d\mu(\theta) \lesssim \frac{b}{\delta^{r/2}(1-2\delta)\sqrt{n\wedge m}}.
\end{aligned} \tag{2.40}$$

Finally, the numerical inequality $|x^r - y^r| \geq y^{r-1}|x-y|$ for all $x, y > 0$ implies

$$\mathbb{E}|D_{nm}| \geq \mathrm{SW}_{r,\delta}^{r-1}(P,Q)\mathbb{E}|\Delta_{nm}| \geq \Gamma^{r-1}\mathbb{E}|\Delta_{nm}|, \tag{2.41}$$

so that $\mathbb{E}|\Delta_{nm}| \lesssim \Gamma^{1-r}b(n\wedge m)^{-1/2}/(\delta^{r/2}(1-2\delta))$, as claimed. □

Proof of Proposition 2(ii). Introduce an i.i.d. sample $X_1', \dots, X_n' \sim P$ independent of $X_1, \dots, X_n$, and let $P_n = \frac{1}{n}\sum_{i=1}^n \delta_{X_i'}$. It follows from convexity of the mapping $(P,Q) \mapsto \mathrm{SW}_{r,\delta}(P,Q)$, similarly as in the proof of Theorem 4.3 of Bobkov and Ledoux (2019), that

$$\mathbb{E}\big[\mathrm{SW}_{r,\delta}^r(P_n,P)\big] \leq \mathbb{E}\big[\mathrm{SW}_{r,\delta}^r(P_n,P_n')\big].$$

The claim now follows from equation (2.40) with $P = Q$, which implies the bound

$$\mathbb{E}\big[\mathrm{SW}_{r,\delta}^r(P_n, P_n')\big] \lesssim \frac{bn^{-1/2}}{\delta^{r/2}(1-2\delta)},$$

so that, $\mathbb{E}\big[\mathrm{SW}_{r,\delta}(P_n, P_n')\big] \lesssim b^{1/r}n^{-1/2r}/(\sqrt{\delta}(1-2\delta)^{1/r})$. □

Proof of Proposition 4(i). The claim is immediate from Proposition 2(i), by the triangle inequality. □

It remains to prove Lemmas 4–6.

2.B.1 Proof of Lemma 4

We shall make use of Bennett's inequality, which we recall as follows using the notation of Pollard (2002).

Lemma 7 (Bennett's Inequality). *Let $Z_1, \dots, Z_n$ be i.i.d. random variables bounded above by 1, and such that $\mathbb{E}[Z_1] = \mu \in \mathbb{R}$, $\mathrm{Var}[Z_1] = v > 0$. Then,*

$$\mathbb{P}\left\{\sum_{i=1}^n (Z_i - \mu) \geq x\right\} \leq \exp\left\{-\frac{x^2}{2W}\psi\left(\frac{x}{W}\right)\right\},$$

for all $x \geq 0$, where $W \geq nv$ is arbitrary, and where for all $t \geq -1$,

$$\psi(t) = \frac{(1+t)\log(1+t) - t}{t^2/2}, \quad \textit{for } t \neq 0, \ \textit{and } \psi(0) = 1.$$

Turning back to the proof, fix $\theta \in \mathbb{S}^{d-1}$. Set $m_\theta = \mathbb{E}(|X^\top\theta|^2)$, and

$$x_\theta = 2\big[|F_\theta^{-1}(\delta/2)| \vee |F_\theta^{-1}(1-\delta/2)| \vee \ell\sqrt{m_\theta/\delta} \vee 1\big],$$

for a constant $\ell > 0$ to be determined below. In particular, for all $x \geq x_\theta$,

$$F_\theta(x) \geq 1 - \delta/2, \quad \text{and} \quad F_\theta(-x) \leq \delta/2.$$

Then, for all $x \geq x_\theta$,

$$\begin{aligned}\mathbb{P}(-F_{\theta,n}^{-1}(\delta) > x) &\leq \mathbb{P}(F_{\theta,n}(-x) > \delta) \\ &\leq \mathbb{P}(F_{\theta,n}(-x) - F_\theta(-x) > \delta/2) \\ &\leq \exp\left\{-\frac{(\delta n)^2}{8W_\theta}\psi\left(\frac{n\delta}{2W_\theta}\right)\right\},\end{aligned}$$

for any given $W_\theta \geq n \operatorname{Var}[I(X^\top\theta \leq -x)] = nF_\theta(-x)(1-F_\theta(-x))$, by Bennett's inequality. Now, from Markov's inequality, one also has

$$F_\theta(-x) = \mathbb{P}(X^\top\theta \leq -x) = \mathbb{P}(-X^\top\theta \geq x) \leq \mathbb{P}(|X^\top\theta| \geq x) \leq m_\theta/x^2,$$

thus, we may take $W_\theta = nm_\theta/x^2$. Furthermore,

$$\begin{aligned}\frac{(\delta n)^2}{8W_\theta}\psi\left(\frac{n\delta}{2W_\theta}\right) &= \frac{(\delta n)^2}{8W_\theta}\frac{(1+\frac{n\delta}{2W_\theta})\log(1+\frac{n\delta}{2W_\theta}) - \frac{n\delta}{2W_\theta}}{(\frac{n\delta}{2W_\theta})^2/2} \\ &= W_\theta\left[\left(1+\frac{n\delta}{2W_\theta}\right)\log\left(1+\frac{n\delta}{2W_\theta}\right) - \frac{n\delta}{2W_\theta}\right] \\ &\geq \frac{nm_\theta}{x^2}\left[\frac{x^2\delta}{2m_\theta}\log\left(1+\frac{x^2\delta}{2m_\theta}\right) - \frac{x^2\delta}{2m_\theta}\right] \\ &\geq n\left[\frac{\delta}{2}\log\left(1+\frac{x^2\delta}{2m_\theta}\right) - \frac{\delta}{2}\right] \\ &\geq \frac{n\delta}{4}\log\left(1+\frac{x^2\delta}{2m_\theta}\right),\end{aligned}$$

where the last inequality holds for all $x \geq x_\theta$ upon choosing the constant $\ell = \sqrt{2(e^2-1)}$ in the definition of x_θ. Therefore, for all such x,

$$\mathbb{P}(-F_{\theta,n}^{-1}(\delta) > x) \leq \left(1+\frac{x^2\delta}{2m_\theta}\right)^{-\frac{n\delta}{4}}$$

Applying a similar argument, we obtain that for all $x \geq x_\theta$,

$$\begin{aligned}\mathbb{P}(F_{\theta,n}^{-1}(1-\delta) > x) &\leq \mathbb{P}(1-F_{\theta,n}(x) > \delta) \\ &\leq \mathbb{P}(F_\theta(x) - F_{\theta,n}(x) > \delta/2) \leq \left(1+\frac{x^2\delta}{2m_\theta}\right)^{-\frac{n\delta}{4}}.\end{aligned}$$

Now notice that

$$|F_{\theta,n}^{-1}(1-\delta)| \leq F_{\theta,n}^{-1}(1-\delta) \vee (-F_{\theta,n}^{-1}(\delta)),$$

thus we arrive at

$$\mathbb{P}(|F_{\theta,n}^{-1}(1-\delta)| \geq x) \leq 2\left(1+\frac{x^2\delta}{2m_\theta}\right)^{-\frac{n\delta}{4}}, \quad x \geq x_\theta.$$

It follows that

$$\begin{aligned}\mathbb{E}|F_{\theta,n}^{-1}(1-\delta)|^{\bar r} &= \bar r\int_0^\infty x^{\bar r-1}\mathbb{P}(|F_{\theta,n}^{-1}(1-\delta)| \geq x)dx \\ &\lesssim x_\theta^{\bar r} + \int_{x_\theta}^\infty x^{\bar r-1}\mathbb{P}(|F_{\theta,n}^{-1}(1-\delta)| \geq x)dx\end{aligned}$$

$$
\begin{aligned}
&\le x_\theta^{\bar r} + 2\int_{x_\theta}^\infty x^{\bar r - 1}\left(1 + \frac{x^2\delta}{2m_\theta}\right)^{-\frac{n\delta}{4}} dx \\
&\le x_\theta^{\bar r} + 2\int_{\sqrt{m_\theta/\delta}}^\infty x^{\bar r - 1}\left(1 + \frac{x^2\delta}{2m_\theta}\right)^{-\frac{n\delta}{4}} dx, \qquad \text{(Since } \ell > 1) \\
&= x_\theta^{\bar r} + 2\left(\frac{m_\theta}{\delta}\right)^{\frac{\bar r}{2}} \int_1^\infty y^{\bar r - 1}\left(1 + \frac{y^2}{2}\right)^{-\frac{n\delta}{4}} dy \\
&\lesssim x_\theta^{\bar r} + \left(\frac{m_\theta}{\delta}\right)^{\frac{\bar r}{2}} \int_1^\infty y^{\bar r - 1 - \frac{n\delta}{2}} dy \\
&= x_\theta^{\bar r} + \left(\frac{m_\theta}{\delta}\right)^{\frac{\bar r}{2}} \frac{1}{\frac{n\delta}{2} - \bar r} \le x_\theta^{\bar r} + \left(\frac{m_\theta}{\delta}\right)^{\frac{\bar r}{2}},
\end{aligned}
$$

where we used the assumptions $\delta \ge 2(r+2)/n$ and $\bar r \le r+1$ on the final line of the above display. Therefore, $\mathbb{E}|F_{\theta,n}^{-1}(1-\delta)|^{\bar r} \lesssim x_\theta^{\bar r} + (m_\theta/\delta)^{\frac{\bar r}{2}}$. Upon repeating the same argument for the δ-quantile, we obtain

$$
\max_{a \in \{\delta, 1-\delta\}} \mathbb{E}|F_{\theta,n}^{-1}(a)|^{\bar r} \lesssim x_\theta^{\bar r} + \left(\frac{m_\theta}{\delta}\right)^{\frac{\bar r}{2}} \lesssim 1 + \max_{a \in \{\frac{\delta}{2}, 1-\frac{\delta}{2}\}} |F_\theta^{-1}(a)|^{\bar r} + \left(\frac{m_\theta}{\delta}\right)^{\frac{\bar r}{2}}.
$$

This proves the claim. □

2.B.2 Proof of Lemma 5

Since $\mathrm{SJ}_{r,\delta/2}(P) < \infty$, it follows from Lemma 2 that F_θ^{-1} is absolutely continuous over $[\delta/2, 1-\delta/2]$ for μ-almost every $\theta \in \mathbb{S}^{d-1}$. We fix any such θ throughout the sequel.

To prove the claim, we shall make use of the following analogue of the relative VC inequality described in Example 2 (Bousquet, Boucheron, and Lugosi, 2003). For any given $\theta \in \mathbb{S}^{d-1}$ and $\epsilon \in (0,1)$, we have

$$
\mathbb{P}\left(|F_{\theta,n}(x) - F_\theta(x)| \le \nu_{\epsilon,n}\sqrt{F_\theta(x)(1-F_\theta(x))},\ \forall x \in \mathbb{R}\right) \ge 1 - \epsilon.
$$

Notice here that the right-hand side of the inequality within the above probability involves the population CDF F_θ rather than $F_{\theta,n}$. Similarly as in Example 2 and Section 2.G, the above bound implies that

$$
\mathbb{P}\Big(F_\theta^{-1}(\gamma_{\epsilon,n}(u)) \le F_{\theta,n}^{-1}(u) \le F_\theta^{-1}(\eta_{\epsilon,n}(u)),\ \forall u \in (0,1)\Big) \ge 1 - \epsilon,
$$

where, for all $u \in (0,1)$,

$$
\begin{aligned}
\gamma_{\epsilon,n}(u) &= \frac{2u + \nu_{\epsilon,n}^2 - \nu_{\epsilon,n}\sqrt{\nu_{\epsilon,n}^2 + 4u(1-u)}}{2(1+\nu_{\epsilon,n}^2)}, \\
\eta_{\epsilon,n}(u) &= \frac{2u + \nu_{\epsilon,n}^2 + \nu_{\epsilon,n}\sqrt{\nu_{\epsilon,n}^2 + 4u(1-u)}}{2(1+\nu_{\epsilon,n}^2)}.
\end{aligned}
\tag{2.42}
$$

These functions are invertible over $[0,1]$, with inverses given by

$$\begin{aligned} \eta_{\epsilon,n}^{-1}(t) &= t - \nu_{\epsilon,n}\sqrt{t(1-t)}, \quad t \in [\eta_{\epsilon,n}(0), \eta_{\epsilon,n}(1)] = \left[\frac{\nu_{\epsilon,n}^2}{1+\nu_{\epsilon,n}^2}, 1\right], \\ \gamma_{\epsilon,n}^{-1}(t) &= t + \nu_{\epsilon,n}\sqrt{t(1-t)}, \quad t \in [\gamma_{\epsilon,n}(0), \gamma_{\epsilon,n}(1)] = \left[0, \frac{1}{1+\nu_{\epsilon,n}^2}\right]. \end{aligned} \tag{2.43}$$

We shall further make use of the following elementary Lemma, proven in Section 2.B.3.

Lemma 8. *Assume $\epsilon \in (0,1)$ is chosen such that $\nu_{\epsilon,n} \leq y_0 := \sqrt{\delta}/4$. Then, for all $u \in [\delta/2, 1-\delta/2]$,*

$$\frac{\gamma_{\epsilon,n}(u)}{u} \geq \frac{1}{2}, \quad \frac{1-\eta_{\epsilon,n}(u)}{1-u} \geq \frac{1}{2}.$$

In particular, for all $x \in [\delta/2, 1-\delta/2]$, and all $y \in [\gamma_{\epsilon,n}(x), \eta_{\epsilon,n}(x)]$, $\frac{x(1-x)}{y(1-y)} \leq \frac{1}{4}$.

By Lemma 8, the inequalities $\gamma_{\epsilon,n}(\delta) \geq \delta/2$ and $\eta_{\epsilon,n}(1-\delta) \leq 1-\delta/2$ hold whenever $\nu_{\epsilon,n} \leq y_0$, and this last inequality is satisfied whenever $\epsilon \geq \epsilon_0 := \frac{2n+1}{16}\exp(-ny_0^2/16)$. For all such ϵ, define the event

$$E_\epsilon \equiv E_\epsilon(\theta) = \Big\{F_\theta^{-1}(\gamma_{\epsilon,n}(u)) \leq F_{\theta,n}^{-1}(u) \leq F_\theta^{-1}(\eta_{\epsilon,n}(u)),\ \forall u \in [\delta, 1-\delta]\Big\}.$$

Over the event E_ϵ, we have

$$\begin{aligned} Z_n(\theta) &= \frac{1}{1-2\delta}\int_\delta^{1-\delta} \left|F_{\theta,n}^{-1}(u) - F_\theta^{-1}(u)\right|^r du \\ &\leq \frac{1}{1-2\delta}\int_\delta^{1-\delta} \left[F_\theta^{-1}(\eta_{\epsilon,n}(u)) - F_\theta^{-1}(\gamma_{\epsilon,n}(u))\right]^r du. \end{aligned}$$

Now, recall that θ was chosen such that F_θ^{-1} is absolutely continuous over $[\delta/2, 1-\delta/2]$, whence

$$F_\theta^{-1}(\eta_{\epsilon,n}(u)) - F_\theta^{-1}(\gamma_{\epsilon,n}(u)) = \int_{\gamma_{\epsilon,n}(u)}^{\eta_{\epsilon,n}(u)} \frac{dt}{p_\theta(F_\theta^{-1}(t))}.$$

Now, since $\nu_{\epsilon,n} \leq y_0$, we have for all $u \in [\delta, 1-\delta]$,

$$\begin{aligned} \eta_{\epsilon,n}(u) - \gamma_{\epsilon,n}(u) &= \frac{\nu_{\epsilon,n}\sqrt{\nu_{\epsilon,n}^2 + 4u(1-u)}}{(1+\nu_{\epsilon,n}^2)} \\ &\leq 2\nu_{\epsilon,n}\sqrt{u(1-u)} + \nu_{\epsilon,n}^2 \leq 3\nu_{\epsilon,n}\sqrt{u(1-u)}. \end{aligned} \tag{2.44}$$

Using Jensen's inequality, we deduce that, over the event E_ϵ,

$$\begin{aligned} &(1-2\delta)Z_n(\theta) \\ &\leq \int_\delta^{1-\delta}\left(\int_{\gamma_{\epsilon,n}(u)}^{\eta_{\epsilon,n}(u)} \frac{1}{p_\theta(F_\theta^{-1}(t))}dt\right)^r du \end{aligned}$$

$$\begin{aligned}
&\leq \int_{\delta}^{1-\delta} (\eta_{\epsilon,n}(u)-\gamma_{\epsilon,n}(u))^{r-1} \int_{\gamma_{\epsilon,n}(u)}^{\eta_{\epsilon,n}(u)} \left(\frac{1}{p_\theta(F_\theta^{-1}(t))}\right)^r dtdu \\
&= \int_{\delta}^{1-\delta} \frac{(\eta_{\epsilon,n}(u)-\gamma_{\epsilon,n}(u))^{r-1}}{(u(1-u))^{r/2}} \int_{\gamma_{\epsilon,n}(u)}^{\eta_{\epsilon,n}(u)} \left(\frac{\sqrt{t(1-t)}}{p_\theta(F_\theta^{-1}(t))}\right)^r \frac{(u(1-u))^{r/2}}{(t(1-t))^{r/2}} dtdu \\
&\lesssim \nu_{\epsilon,n}^{r-1} \int_{\delta}^{1-\delta} \frac{1}{\sqrt{u(1-u)}} \int_{\gamma_{\epsilon,n}(u)}^{\eta_{\epsilon,n}(u)} \left(\frac{\sqrt{t(1-t)}}{p_\theta(F_\theta^{-1}(t))}\right)^r dtdu \ \text{(By (2.44), Lem. 8)} \\
&\leq \nu_{\epsilon,n}^{r-1} \int_{\delta/2}^{1-\delta/2} \left(\int_{\eta_{\epsilon,n}^{-1}(t)}^{\gamma_{\epsilon,n}^{-1}(t)} \frac{1}{\sqrt{u(1-u)}} du\right) \left(\frac{\sqrt{t(1-t)}}{p_\theta(F_\theta^{-1}(t))}\right)^r dt,
\end{aligned}$$

where we interchanged the order of integration, and used the fact that $\delta/2 \leq \gamma_{\epsilon,n}(u) \leq \eta_{\epsilon,n}(u) \leq 1-\delta/2$ for all $u \in [\delta, 1-\delta]$. We further have

$$\int_{\eta_{\epsilon,n}^{-1}(t)}^{\gamma_{\epsilon,n}^{-1}(t)} \frac{1}{\sqrt{u(1-u)}} du \leq \frac{\gamma_{\epsilon,n}^{-1}(t)-\eta_{\epsilon,n}^{-1}(t)}{\sqrt{u_t^*(1-u_t^*)}}, \quad \text{for } u_t^* \in \underset{\eta_{\epsilon,n}^{-1}(t) \leq u \leq \gamma_{\epsilon,n}^{-1}(t)}{\operatorname{argmin}} \sqrt{u(1-u)}.$$

By again applying Lemma 8, and equation (2.43), we obtain

$$\frac{\gamma_{\epsilon,n}^{-1}(t)-\eta_{\epsilon,n}^{-1}(t)}{\sqrt{u_t^*(1-u_t^*)}} \leq \frac{\nu_{\epsilon,n}\sqrt{t(1-t)}}{\sqrt{u_t^*(1-u_t^*)}} \leq \frac{1}{2}\nu_{\epsilon,n}.$$

We have thus shown

$$Z_n(\theta) \lesssim \frac{\nu_{\epsilon,n}^r}{1-2\delta} \int_{\delta/2}^{1-\delta/2} \left(\frac{\sqrt{t(1-t)}}{p_\theta(F_\theta^{-1}(t))}\right)^r dt = \nu_{\epsilon,n}^r J_{r,\delta/2}(P_\theta).$$

Thus, setting $\epsilon = \frac{2n+1}{16} e^{-\frac{ny^2}{16}}$ for any $y \in (0, y_0]$, we have

$$\mathbb{P}\Big(Z_n(\theta) \geq C_r y^r J_{r,\delta/2}(P_\theta)\Big) \leq \epsilon,$$

for a universal constant $C_r > 0$ depending only on r. □

2.B.3 Proof of Lemma 8

For all $u \in [\delta/2, 1-\delta/2]$, and all ϵ such that $\nu_{\epsilon,n}^2 \leq \delta/16$, we have

$$\begin{aligned}
\frac{\gamma_{\epsilon,n}(u)}{u} &\geq \frac{1}{1+\nu_{\epsilon,n}^2} - \frac{\nu_{\epsilon,n}\sqrt{\nu_{\epsilon,n}^2+4u(1-u)}-\nu_{\epsilon,n}^2}{2u(1+\nu_{\epsilon,n}^2)} \\
&\geq \frac{1}{1+\nu_{\epsilon,n}^2} - \frac{\nu_{\epsilon,n}\sqrt{1-u}}{\sqrt{u}(1+\nu_{\epsilon,n}^2)} \\
&\geq \frac{1}{1+\nu_{\epsilon,n}^2} - \frac{\sqrt{2}\sqrt{\delta}\sqrt{1-u}}{4\sqrt{\delta}(1+\nu_{\epsilon,n}^2)}
\end{aligned}$$

$$\begin{aligned}
&\geq \frac{1}{1+\nu_{\epsilon,n}^2} - \frac{1}{2\sqrt{2}(1+\nu_{\epsilon,n}^2)} \\
&\geq \frac{2\sqrt{2}-1}{2\sqrt{2}(1+\nu_{\epsilon,n}^2)} \geq \frac{2\sqrt{2}-1}{2\sqrt{2}(1+1/16)} \geq 1/2.
\end{aligned}$$

Similarly,

$$\begin{aligned}
\frac{1-\eta_{\epsilon,n}(u)}{1-u} &\geq \frac{1}{1+\nu_{\epsilon,n}^2} - \frac{\nu_{\epsilon,n}^2 + \nu_{\epsilon,n}\sqrt{\nu_{\epsilon,n}^2 + 4u(1-u)}}{2(1-u)(1+\nu_{\epsilon,n}^2)} \\
&\geq \frac{1}{1+\nu_{\epsilon,n}^2} - \frac{\nu_{\epsilon,n}^2 + \nu_{\epsilon,n}\sqrt{u(1-u)}}{(1-u)(1+\nu_{\epsilon,n}^2)} \\
&= \frac{1-\nu_{\epsilon,n}^2/(1-u)}{1+\nu_{\epsilon,n}^2} - \frac{\nu_{\epsilon,n}\sqrt{u}}{\sqrt{1-u}(1+\nu_{\epsilon,n}^2)} \\
&\geq \frac{1-\nu_{\epsilon,n}^2/\delta}{1+\nu_{\epsilon,n}^2} - \frac{\sqrt{2}\nu_{\epsilon,n}}{\sqrt{\delta}(1+\nu_{\epsilon,n}^2)} \\
&\geq \frac{1-1/16}{1+\nu_{\epsilon,n}^2} - \frac{1}{2\sqrt{2}(1+\nu_{\epsilon,n}^2)} \\
&\geq \frac{2\sqrt{2}(1-1/16)-1}{2\sqrt{2}(1+1/16)} \geq 1/2.
\end{aligned}$$

In particular, for all $x \in [\delta/2, 1-\delta/2]$, and all $y \in [\gamma_{\epsilon,n}(x), \eta_{\epsilon,n}(x)]$,

$$\frac{x(1-x)}{y(1-y)} \leq \frac{x}{\gamma_{\epsilon,n}(x)} \frac{1-x}{1-\eta_{\epsilon,n}(x)} \leq \frac{1}{2}.$$

The claim follows. □

2.B.4 Proof of Lemma 6

When $r > 1$, we have,

$$\begin{aligned}
&(1-2\delta)\mathrm{SW}_{r,\delta}^r(P_n, Q_m) \\
&\quad = \int_{\mathbb{S}^{d-1}} \int_\delta^{1-\delta} \left|F_{\theta,n}^{-1}(u) - G_{\theta,m}^{-1}(u)\right|^r du d\mu(\theta) \\
&\quad = \int_{\mathbb{S}^{d-1}} \int_\delta^{1-\delta} \Big\{ |F_\theta^{-1}(u) - G_\theta^{-1}(u)|^r + r\,\mathrm{sgn}(\widetilde{F}_{\theta,n}^{-1}(u) - \widetilde{G}_{\theta,m}^{-1}(u)) \\
&\times |\widetilde{F}_{\theta,n}^{-1}(u) - \widetilde{G}_{\theta,m}^{-1}(u)|^{r-1} \big\{ (F_{\theta,n}^{-1}(u) - F_\theta^{-1}(u)) - (G_{\theta,m}^{-1}(u) - G_\theta^{-1}(u)) \big\} \Big\} du d\mu(\theta),
\end{aligned}$$

by a Taylor expansion of the map $(x,y) \mapsto |x-y|^r$ about $\big(F_\theta^{-1}(u), G_\theta^{-1}(u)\big)$, where $\widetilde{F}_{\theta,n}^{-1}(u)$ (resp. $\widetilde{G}_{\theta,m}^{-1}(u)$) is a real number on the line joining $F_\theta^{-1}(u)$ and $F_{\theta,n}^{-1}(u)$ (resp. $G_\theta^{-1}(u)$ and

$G_{\theta,m}^{-1}(u)$). We then have

$$\begin{aligned}
|D_{nm}| &\leq \frac{r}{1-2\delta}\int_{\mathbb{S}^{d-1}}\int_{\delta}^{1-\delta}|\widetilde{F}_{\theta,n}^{-1}(u)-\widetilde{G}_{\theta,m}^{-1}(u)|^{r-1} \\
&\qquad\qquad \times\Big[|F_{\theta,n}^{-1}(u)-F_{\theta}^{-1}(u)|+|G_{\theta,m}^{-1}(u)-G_{\theta}^{-1}(u)|\Big]dud\mu(\theta) \\
&\leq \frac{r}{1-2\delta}\int_{\mathbb{S}^{d-1}}(b_\theta-a_\theta)^{r-1} \\
&\qquad\qquad \times\int_{\delta}^{1-\delta}\Big[|F_{\theta,n}^{-1}(u)-F_{\theta}^{-1}(u)|+|G_{\theta,m}^{-1}(u)-G_{\theta}^{-1}(u)|\Big]dud\mu(\theta) \\
&= \frac{r}{1-2\delta}\int_{\mathbb{S}^{d-1}}Z_{nm}(\theta)d\mu(\theta).
\end{aligned}$$

This proves the first claim when $r > 1$, and the same conclusion holds trivially when $r = 1$ by the triangle inequality. To prove the second claim, given $\theta \in \mathbb{S}^{d-1}$, define the following event for any $t \in (0, \delta/2]$,

$$\begin{aligned}
E_t \equiv E_t(\theta) = \Big\{F_{\theta,n}^{-1}(u-t) \leq F_{\theta}^{-1}(u) \leq F_{\theta,n}^{-1}(u+t),\ \ \forall u\in[\delta,1-\delta]\Big\} \\
\cap\Big\{G_{\theta,m}^{-1}(u-t) \leq G_{\theta}^{-1}(u) \leq G_{\theta,m}^{-1}(u+t),\ \ \forall u\in[\delta,1-\delta]\Big\}.
\end{aligned}$$

A union bound together with the Dvoretzky-Kiefer-Wolfowitz inequality (Example 1) implies that, for all $t \in (0,\delta/2]$, $\mathbb{P}(E_t) \geq 1-4\exp(-2(n\wedge m)t^2)$. Now, for all such t, the following inequalities hold over E_t,

$$\begin{aligned}
&\int_{\delta}^{1-\delta}|F_{\theta,n}^{-1}(u)-F_{\theta}^{-1}(u)|du \\
&\quad\leq \int_{\delta}^{1-\delta}\big[F_{\theta}^{-1}(u+t)-F_{\theta}^{-1}(u-t)\big]du \\
&\quad= \int_{\delta-t}^{1-\delta-t}F_{\theta}^{-1}(u)du-\int_{\delta+t}^{1-\delta+t}F_{\theta}^{-1}(u)du \\
&\quad= \int_{\delta-t}^{\delta+t}F_{\theta}^{-1}(u)du+\int_{1-\delta-t}^{1-\delta+t}F_{\theta}^{-1}(u)du \\
&\quad\leq 2t\big[|F_{\theta}^{-1}(\delta-t)|+|F_{\theta}^{-1}(\delta+t)|+|F_{\theta}^{-1}(1-\delta+t)|+|F_{\theta}^{-1}(1-\delta-t)|\big] \\
&\quad\leq 2t\big[|F_{\theta}^{-1}(\delta/2)|+|F_{\theta}^{-1}(1-\delta/2)|\big].
\end{aligned}$$

Over E_t, we also have for all $t \in (0,\delta/2]$,

$$\int_{\delta}^{1-\delta}|G_{\theta,n}^{-1}(u)-G_{\theta}^{-1}(u)|du \leq 2t\big[|G_{\theta}^{-1}(\delta/2)|+|G_{\theta}^{-1}(1-\delta/2)|\big],$$

and,

$$a_\theta \geq F_{\theta}^{-1}(\delta-t)\wedge G_{\theta}^{-1}(\delta-t) \geq F_{\theta}^{-1}(\delta/2)\wedge G_{\theta}^{-1}(\delta/2),$$

$$b_\theta \leq F_\theta^{-1}\big(1-\delta+t\big) \vee G_\theta^{-1}(1-\delta+t) \leq F_\theta^{-1}(1-\delta/2) \vee G_\theta^{-1}(1-\delta/2).$$

Combining these facts, we deduce that for a universal constant $A > 0$, we have with probability at least $1 - 4\exp(-2(n \wedge m)t^2)$,

$$|Z_{nm}(\theta)| \leq AtM^r(\theta),$$

as was to be shown. □

2.C Proof of Theorem 5

Throughout the proof, KL denotes the Kullback-Leibler divergence, and χ^2 denotes the χ^2-divergence. In view of the identity $W_{r,\delta}(P,Q) = W_r(P^\delta, Q^\delta)$ stated in Section 2.2.1, and its natural analogue for the Sliced Wasserstein distance, together with the fact that all distributions considered below are compactly supported, there will be no loss of generality in assuming $\delta = 0$ in what follows.

At a high-level, our general approach is to carefully construct two pairs of distributions $(P_0, Q_0), (P_1, Q_1) \in \mathcal{O}(\Gamma; s_1, s_2)$ such that the corresponding product measures $(P_0^{\otimes n} \otimes Q_0^{\otimes m})$ and $(P_1^{\otimes n} \otimes Q_1^{\otimes m})$ are close in the KL distance, but such that $\mathrm{SW}_r(P_0, Q_0)$ and $\mathrm{SW}_r(P_1, Q_1)$ are sufficiently different. In particular, if we can show that $\mathrm{KL}\big(P_0^{\otimes n} \otimes Q_0^{\otimes m}, P_1^{\otimes n} \otimes Q_1^{\otimes m}\big) \leq \zeta < \infty$, then via an application of Le Cam's inequality (see for instance, Theorem 2.2 of Tsybakov (2008)), we obtain the minimax lower bound that,

$$\mathcal{R}_{nm}(\mathcal{O}(\Gamma; s_1, s_2); r) \geq c_\zeta |\mathrm{SW}_r(P_0, Q_0) - \mathrm{SW}_r(P_1, Q_1)|, \tag{2.45}$$

where $c_\zeta > 0$ is a constant depending only on ζ. We will use four separate constructions to handle various cases of the Theorem.

Let $\epsilon_n = k_r n^{-1/2}$, for a constant $k_r \in (0,1)$, possibly depending on r, to be determined below. We use the following pairs of distributions.

- **Construction 1.** For a vector $A = (a, 0, \dots, 0) \in \mathbb{R}^d$, and for $g > 0$, we define:

$$P_{01} = \frac{1}{2}\delta_0 + \frac{1}{2}\delta_A, \qquad Q_{01} = \frac{1}{2}\delta_{gA} + \frac{1}{2}\delta_{(1+g)A}$$
$$P_{11} = \left(\frac{1}{2} + \epsilon_n\right)\delta_0 + \left(\frac{1}{2} - \epsilon_n\right)\delta_A, \quad Q_{11} = \frac{1}{2}\delta_{gA} + \frac{1}{2}\delta_{(1+g)A}.$$

- **Construction 2.** For $\gamma_2, \Delta > 0$ to be chosen in the sequel we let $P_{02}, P_{12}, Q_{02}, Q_{12} \in \mathcal{P}(\mathbb{R}^d)$ be the probability distributions of random vectors of the form $(X, 0, \dots, 0) \in \mathbb{R}^d$, with X respectively distributed according to the distributions

$$P_{02}^{(1)} = U\left(0, \gamma_2^{1/r}\right), \qquad Q_{02}^{(1)} = U\left(\Delta\gamma_2^{\frac{1}{r}}, (1+\Delta)\gamma_2^{\frac{1}{r}}\right)$$
$$P_{12}^{(1)} = \frac{1+\epsilon_n}{2} U\Big(0, \frac{\gamma_2^{\frac{1}{r}}}{2}\Big) + \frac{1-\epsilon_n}{2} U\Big(\frac{\gamma_2^{\frac{1}{r}}}{2}, \gamma_2^{\frac{1}{r}}\Big), \quad Q_{12}^{(1)} = U\left(\Delta\gamma_2^{\frac{1}{r}}, (1+\Delta)\gamma_2^{\frac{1}{r}}\right).$$

- **Construction 3.** For $0 < s_1 \le s_2$ we let P_{03}, P_{13}, Q_{03}, $Q_{13} \in \mathcal{P}(\mathbb{R}^d)$ be the probability distributions of random vectors of the form $(X, 0, \dots, 0) \in \mathbb{R}^d$, with X respectively distributed according to the distributions

$$\begin{aligned} P_{03}^{(1)} &= U\left(0, s_1^{1/r}\right), & Q_{03}^{(1)} &= U\left(0, s_2^{1/r}\right), \\ P_{13}^{(1)} &= U\left(0, s_1^{1/r}\right), & Q_{13}^{(1)} &= (1-\epsilon_m) U\left(0, s_2^{1/r}\right) + \epsilon_m \delta_{s_2^{1/r}}. \end{aligned}$$

- **Construction 4.** For $0 < s_2 \le s_1$ we let $P_{04}, P_{14}, Q_{04}, Q_{14} \in \mathcal{P}_r(\mathbb{R}^d)$ be the probability distributions of random vectors of the form $(X, 0, \dots, 0) \in \mathbb{R}^d$, with X respectively distributed according to the distributions

$$\begin{aligned} P_{04}^{(1)} &= U\left(0, s_1^{1/r}\right), & Q_{04}^{(1)} &= U\left(0, s_2^{1/r}\right), \\ P_{14}^{(1)} &= (1-\epsilon_n) U\left(0, s_1^{1/r}\right) + \epsilon_n \delta_{s_1^{1/r}}, & Q_{14}^{(1)} &= U\left(0, s_2^{1/r}\right). \end{aligned}$$

Construction 1 uses pairs of distributions with infinite SJ_r, while Constructions 2-4 use pairs of distributions with finite SJ_r. To compactly state our next result we define several terms,

$$\begin{aligned} t_r &:= \left(\int_{\mathbb{S}^{d-1}} |\theta_1|^r d\mu(\theta) \right)^{\frac{1}{r}}, \\ \beta &:= (s_2/s_1)^{1/r}, & \bar{\beta} &:= 1/\beta, \\ \Delta_\beta &:= \beta - 1, & \Delta_{\bar{\beta}} &:= \bar{\beta} - 1. \end{aligned}$$

With these definitions in place the following technical lemma describes the main features of our constructions.

Lemma 9. *There exists a choice of constant $k_r \in (0,1)$ for which the following statements hold.*

- **Construction 1.** *Let $g := \Gamma / \left(\int_{\mathbb{S}^{d-1}} |A^\top \theta|^r d\mu(\theta) \right)^{1/r}$. Then, there exists a constant $c_1 > 0$, possibly depending on r, such that*

$$\begin{aligned} \mathrm{SW}_r^r(P_{01}, Q_{01}) &= \Gamma^r, \\ \mathrm{SW}_r^r(P_{11}, Q_{11}) &\ge \Gamma^r + c_1 \epsilon_n. \end{aligned}$$

Furthermore, there exists a choice of the vector A for which $P_{01}, Q_{01}, P_{11}, Q_{11} \in \mathcal{O}(\Gamma; \infty, \infty)$.

- **Construction 2.** *There exists a constant $c_2 > 0$, possibly depending on r, such that*

$$\begin{aligned} \mathrm{SW}_r^r(P_{02}, Q_{02}) &= \gamma_2 (t_r \Delta)^r, \\ \mathrm{SW}_r^r(P_{12}, Q_{12}) &\ge \gamma_2 t_r^r \Big\{ \Delta^r + c_2 \Delta^{r-1} \epsilon_n \Big\}. \end{aligned}$$

Furthermore, $P_{02}, Q_{02}, P_{12}, Q_{12} \in \mathcal{O}(0; \gamma_2, \gamma_2)$.

- **Construction 3.** *Assume that* $\bar\beta \in (0,1]$. *Then,*
$$\mathrm{SW}_r^r(P_{03}, Q_{03}) = \frac{s_2 t_r^r |\Delta_{\bar\beta}|^r}{r+1},$$
$$\mathrm{SW}_r^r(P_{13}, Q_{13}) \geq \frac{t_r^r s_2}{r+1}\Big\{|\Delta_{\bar\beta}|^r + r|\Delta_{\bar\beta}|^{r-1}\epsilon_m\Big\}.$$
Furthermore, $P_{03}, Q_{03}, P_{13}, Q_{13} \in \mathcal{O}(0; s_1, s_2)$.

- **Construction 4.** *Assume that* $\beta \in (0,1]$. *Then,*
$$\mathrm{SW}_r^r(P_{04}, Q_{04}) = \frac{s_1 t_r^r |\Delta_{\beta}|^r}{r+1},$$
$$\mathrm{SW}_r^r(P_{14}, Q_{14}) \geq \frac{t_r^r s_1}{r+1}\Big\{|\Delta_{\beta}|^r + r|\Delta_{\beta}|^{r-1}\epsilon_n\Big\}.$$
Furthermore, $P_{04}, Q_{04}, P_{14}, Q_{14} \in \mathcal{O}(0; s_1, s_2)$.

In each case, for some fixed universal constant $\zeta > 0$ *we have that,*
$$\mathrm{KL}\big(P_{0i}^{\otimes n} \otimes Q_{0i}^{\otimes m}, P_{1i}^{\otimes n} \otimes Q_{1i}^{\otimes m}\big) \leq \zeta < \infty, \quad i = 1, 2, 3, 4.$$

Taking this result as given, we can now complete the proof of the Theorem. Using Construction 1 with $\Gamma = 0$, we obtain from equation (2.45) that
$$\mathcal{R}_{nm}(\mathcal{O}(0; \infty, \infty); r) \geq c_\zeta |\mathrm{SW}_r(P_{01}, Q_{01}) - \mathrm{SW}_r(P_{11}, Q_{11})| \gtrsim \epsilon_n^{1/r} \asymp n^{-1/2r}.$$
Reversing the roles of n and m we obtain the first claim of part (i) of the Theorem. Choosing Γ to be a strictly positive constant, we instead obtain
$$\mathcal{R}_{nm}(\mathcal{O}(\Gamma; \infty, \infty); r) \gtrsim \Gamma \left|1 - \left(1 + \frac{c_1 \epsilon_n}{\Gamma^r}\right)^{\frac{1}{r}}\right| = \frac{c_1 k_r n^{-1/2}}{r\Gamma^{r-1}}(1 + o(1)),$$
by a first-order Taylor expansion of the map $x \mapsto (1+x)^{\frac{1}{r}}$. The fact that Γ is bounded away from zero then implies $\mathcal{R}_{nm}(\mathcal{O}(\Gamma; \infty, \infty); r) \gtrsim n^{-1/2}$ which proves part (ii) of the theorem, again upon reversing the roles of n and m. It thus only remains to prove the second claim of part (i).

Without loss of generality we assume that $n \leq m$ in the remainder of the proof, noting that, as above, we may always reverse the roles of n and m and repeat our constructions. We consider four cases.

Case 1: $-1 \leq \Delta_\beta \leq -\epsilon_n$. In this case, the condition $\Delta_\beta \leq 0$ implies $s_1 \geq s_2$. Since $n \leq m$, it therefore suffices to prove $\mathcal{R}_{nm}(\mathcal{O}(0, s_1, s_2); r) \gtrsim s_1^{1/r}\epsilon_n$. Furthermore, since $\beta \leq 1$, we may invoke Construction 4 to obtain
$$|\mathrm{SW}_r(P_{04}, Q_{04}) - \mathrm{SW}_r(P_{14}, Q_{14})| \geq \frac{s_1^{1/r} t_r |\Delta_\beta|}{(r+1)^{1/r}}\left[\left(1 + \frac{r\epsilon_n}{|\Delta_\beta|}\right)^{\frac{1}{r}} - 1\right] \asymp s_1^{1/r}\epsilon_n,$$
where we have used the assumption $|\Delta_\beta| \geq \epsilon_n$ in the last order assesment of the above display. This fact together with equation (2.45) yields the desired lower bound for Case 1.

Case 2: $-\epsilon_n < \Delta_\beta \leq 0$. The inequality $s_1 \geq s_2$ continues to hold, thus it suffices to prove $\mathcal{R}_{nm}(\mathcal{O}(0; s_1, s_2); r) \gtrsim s_1^{1/r}\epsilon_n$. Notice further that

$$s_2^{1/r}\epsilon_n = s_1^{1/r}\beta\epsilon_n > s_1^{1/r}(1-\epsilon_n)\epsilon_n = s_1^{1/r}\epsilon_n(1+o(1)).$$

It will therefore suffice to prove $\mathcal{R}_{nm}(\mathcal{O}(0; s_1, s_2); r) \gtrsim s_2^{1/r}\epsilon_n$. We use Construction 2, and choose $\gamma_2 = s_2$, and $\Delta \in (0, 1]$ to be a constant larger than ϵ_n. We observe that all distributions have SJ_r value at most $s_2 = \min\{s_1, s_2\}$. Furthermore,

$$\left|\mathrm{SW}_r(P_{02}, Q_{02}) - \mathrm{SW}_r(P_{12}, Q_{12})\right| \geq s_2^{1/r} t_r \Delta \left\{ \left(1 + \frac{c_2\epsilon_n}{\Delta}\right)^{\frac{1}{r}} - 1 \right\}.$$

Since $\Delta \geq \epsilon_n$, it is a straightforward observation that the factor in braces of the above display is of order ϵ_n, thus we have shown

$$\left|\mathrm{SW}_r(P_{02}, Q_{02}) - \mathrm{SW}_r(P_{12}, Q_{12})\right| \gtrsim s_2^{1/r}\epsilon_n,$$

and this together with equation (2.45) yields the desired lower bound for Case 2.

Case 3: $-1 \leq \Delta_{\bar\beta} \leq -\epsilon_m$ and $s_1^{1/r}\epsilon_n \leq s_2^{1/r}\epsilon_m$. In this case, it suffices to prove that $\mathcal{R}_{nm}(\mathcal{O}(0, s_1, s_2); r) \gtrsim s_2^{1/r}\epsilon_m$. Notice that $\bar\beta \leq 1$, hence we may use Construction 3 to obtain

$$|\mathrm{SW}_r(P_{03}, Q_{03}) - \mathrm{SW}_r(P_{13}, Q_{13})| \geq \frac{s_2^{1/r} t_r |\Delta_{\bar\beta}|}{(r+1)^{1/r}} \left[\left(1 + \frac{\epsilon_m}{|\Delta_{\bar\beta}|}\right)^{\frac{1}{r}} - 1 \right] \asymp s_2^{1/r}\epsilon_m,$$

where we have used the assumption $|\Delta_{\bar\beta}| \geq \epsilon_m$ in the last order assesment of the above display. This fact together with equation (2.45) yields the desired lower bound for Case 1.

Case 4: $-\epsilon_m < \Delta_{\bar\beta} < 0$ or $s_1^{1/r}\epsilon_n > s_2^{1/r}\epsilon_m$. Notice that if the condition $\Delta_{\bar\beta} > -\epsilon_m$ is satisfied, it implies

$$s_1^{1/r}\epsilon_n = s_2^{1/r}\bar\beta\epsilon_n > (1-\epsilon_m)\epsilon_n s_2^{1/r} \geq (1-\epsilon_m)\epsilon_m s_2^{1/r} \gtrsim \epsilon_m s_2^{1/r}.$$

For this case, it will thus suffice to prove $\mathcal{R}_{nm}(\mathcal{O}(0; s_1, s_2); r) \gtrsim s_1^{1/r}\epsilon_n$. Since $\Delta_{\bar\beta} \leq 0$, we observe that all distributions have SJ_r value at most $s_1 = \min\{s_1, s_2\}$. Invoking Construction 2 with $\gamma_2 = s_1$, the remainder of the argument follows similarly as in Case 2.

It remains to establish Lemma 9 and we turn our attention to this now.

2.C.1 Proof of Lemma 9

Bounding the KL divergence in each case is straightforward. We observe that for each $1 \leq i \leq 4$,

$$\mathrm{KL}(P_{0i}^{\otimes n} \otimes Q_{0i}^{\otimes m}, (P_{1i}^{\otimes n} \otimes Q_{1i}^{\otimes m})) = n\mathrm{KL}(P_{0i}, P_{1i}) + m\mathrm{KL}(Q_{0i}, Q_{1i})$$

$$\leq n\chi^2(P_{0i}, P_{1i}) + m\chi^2(Q_{0i}, Q_{1i}).$$

The χ^2 divergences in each construction can be computed in closed form. Doing so yields the bounds:

$$\begin{aligned}
\mathrm{KL}\big(P_{0i}^{\otimes n} \otimes Q_{0i}^{\otimes m}, P_{1i}^{\otimes n} \otimes Q_{1i}^{\otimes m}\big) &\lesssim n\epsilon_n^2, \quad i = 1, 2, 4 \\
\mathrm{KL}\big(P_{03}^{\otimes n} \otimes Q_{03}^{\otimes m}, P_{13}^{\otimes n} \otimes Q_{13}^{\otimes m}\big) &\lesssim m\epsilon_m^2.
\end{aligned}$$

Together with the definition of ϵ_n, we obtain the desired bounds on the KL divergence.

As a second preliminary, let us verify that, for appropriate choice of various parameters, the distributions we construct have appropriately bounded moments, and belong to the class $\mathcal{K}_r(b)$ defined in equation (2.9). Notice first that the distributions $P_{01}, Q_{01}, P_{11}, Q_{11}$ have support with diameter bounded above by $(1 + G)a$. Choosing a (possibly depending on G and hence Γ) such that this expression is bounded above by $b^{1/r}$ ensures $P_{01}, Q_{01}, P_{11}, Q_{11} \in \mathcal{K}_r(b)$. We are guaranteed that such a choice exists by using the assumption $\Gamma^r \leq c_r b$, which ensures that Γ cannot be too large.

Furthermore, the distributions P_{ij}, Q_{ij} for $i = 2, 3, 4$ and $j = 0, 1$ have supports with diameter bounded above by $s(1 + \Delta) \leq 2s$. The assumption $b \geq (2s)^{1/r}$ therefore guarantees $P_{ij}, Q_{ij} \in \mathcal{K}_r(b)$ for $i = 2, 3, 4$ and $j = 0, 1$.

We now consider each construction in turn, establishing the remaining claims. As a preliminary technical result, it will be useful to study the Wasserstein distance between several pairs of univariate distributions.

Lemma 10. *1. Let $\Delta \geq \epsilon > 0$, and define the distributions*

$$\nu = \frac{1+\epsilon}{2} U(0, 1/2) + \frac{1-\epsilon}{2} U(1/2, 1),$$

and $\rho = U(\Delta, 1 + \Delta)$. Then,

$$W_r^r(\nu, \rho) \geq \Delta^r + \frac{r}{4}\epsilon\Delta^{r-1}.$$

2. Given $\xi \in (0, 1]$, $\Delta_\xi = \xi - 1$, define for all $\epsilon \in (0, 1]$,

$$\nu = U(0, \xi), \quad \rho = (1 - \epsilon)U(0, 1) + \epsilon\delta_1.$$

Then,

$$W_r(\nu, \rho) \geq \frac{1}{r+1}\Big[|\Delta_\xi|^r + r\epsilon|\Delta_\xi|^{r-1}\Big]$$

We prove this result in Section 2.C.1.1. Taking this result as given, we can now compute the various Sliced Wasserstein distances and SJ_r evaluations.

Computing the Sliced Wasserstein distances.

- **Construction 1.** For any $\theta \in \mathbb{S}^{d-1}$, let $F^{-1}_{01,\theta}$, $F^{-1}_{11,\theta}$, $G^{-1}_{01,\theta}$ and $G^{-1}_{11,\theta}$ denote the respective quantile functions of $\pi_\theta\#P_{01}$, $\pi_\theta\#P_{11}$, $\pi_\theta\#Q_{01}$, $\pi_\theta\#Q_{11}$. We have

$$F^{-1}_{01,\theta}(u) = \begin{cases} 0 \wedge A^\top\theta, & u \in (0,1/2) \\ 0 \vee A^\top\theta, & u \in [1/2,1), \end{cases}$$

$$F^{-1}_{11,\theta}(u) = \begin{cases} 0 \wedge A^\top\theta, & u \in (0,1/2+\epsilon_n) \\ 0 \vee A^\top\theta, & u \in [1/2+\epsilon_n,1), \end{cases}$$

$$G^{-1}_{01,\theta}(u) = G^{-1}_{11,\theta} = \begin{cases} gA^\top\theta \wedge (1+g)A^\top\theta, & u \in (0,1/2) \\ gA^\top\theta \vee (1+g)A^\top\theta, & u \in [1/2,1). \end{cases}$$

Therefore,

$$\begin{aligned}
&\mathrm{SW}_r^r(P_{01},Q_{01}) \\
&= \int_{\mathbb{S}^{d-1}}\int_0^1 \left|F^{-1}_{01,\theta}(u) - G^{-1}_{01,\theta}(u)\right|^r du d\mu(\theta) \\
&= \frac{1}{2}\int_{\{\theta\in\mathbb{S}^{d-1}:A^\top\theta\geq 0\}} |gA^\top\theta|^r d\mu(\theta) \\
&\quad + \frac{1}{2}\int_{\{\theta\in\mathbb{S}^{d-1}:A^\top\theta< 0\}} |A^\top\theta - (1+g)A^\top\theta|^r d\mu(\theta) \\
&= g^r\int_{\mathbb{S}^{d-1}} |A^\top\theta|^r d\mu(\theta) = \Gamma^r.
\end{aligned}$$

Furthermore,

$$\begin{aligned}
&\mathrm{SW}_r^r(P_{11},Q_{11}) \\
&= \int_{\mathbb{S}^{d-1}}\int_0^1 \left|F^{-1}_{11,\theta}(u) - G^{-1}_{11,\theta}(u)\right|^r du d\mu(\theta) \\
&= \int_{\mathbb{S}^{d-1}}\Bigg(\int_0^{1/2} |0\wedge A^\top\theta - gA^\top\theta\wedge(1+g)A^\top\theta|^r du \\
&\qquad + \int_{1/2}^{1/2+\epsilon_n} |0\wedge A^\top\theta - gA^\top\theta\vee(1+g)A^\top\theta|^r du \\
&\qquad + \int_{1/2+\epsilon_n}^{1} |0\vee A^\top\theta - gA^\top\theta\vee(1+g)\theta^\top A|^r du\Bigg) d\mu(\theta) \\
&= (1-\epsilon_n)g^r\int_{\mathbb{S}^{d-1}} |A^\top\theta|^r d\mu(\theta) \\
&\qquad + \epsilon_n\int_{\mathbb{S}^{d-1}} |0\wedge A^\top\theta - gA^\top\theta\vee(1+g)A^\top\theta|^r d\mu(\theta) \\
&= (1-\epsilon_n)g^r I_r^r + \epsilon_n\int_{\mathbb{S}^{d-1}} |0\wedge A^\top\theta - gA^\top\theta\vee(1+g)A^\top\theta|^r d\mu(\theta)
\end{aligned}$$

$$= \Gamma^r + c_1\epsilon_n,$$

for a positive constant $c_1 > 0$. It follows that $\mathrm{SW}_r(P_{11}, Q_{11}) \geq \mathrm{SW}_r(P_{01}, Q_{01}) \geq \Gamma$, thus $(P_{01}, Q_{01}), (P_{11}, Q_{11}) \in \mathcal{O}(\Gamma; \infty, \infty)$, and

$$\left|\mathrm{SW}_r(P_{01}, Q_{01}) - \mathrm{SW}_r(P_{11}, Q_{11})\right| = \left|\Gamma - \left(\Gamma^r + c_1\epsilon_n\right)^{\frac{1}{r}}\right|.$$

- **Construction 2.** We use the first part of Lemma 10, and let $\nu = \frac{1+\epsilon_n}{2}U(0, 1/2) + \frac{1-\epsilon_n}{2}U(1/2, 1)$, and $\rho = U(\Delta, 1+\Delta)$. Notice that if $X \sim \nu$, then $\gamma_2^{1/r}X \sim P_{12}^{(1)}$, and if $Y \sim \rho$, then $\gamma_2^{1/r}Y \sim Q_{02}^{(1)}$. Therefore, by Proposition 7.16 of Villani (2003), $W_r(\pi_{\theta\#}P_{12}, \pi_{\theta\#}Q_{12}) = |\theta_1|\gamma_2^{1/r}W_r(\nu, \rho)$. Thus,

  $$\begin{aligned}\mathrm{SW}_r^r(P_{12}, Q_{12}) &= \int_{\mathbb{S}^{d-1}} W_r^r(\pi_{\theta\#}P_{12}, \pi_{\theta\#}Q_{12})d\mu(\theta) \\ &= \int_{\mathbb{S}^{d-1}} |\theta_1|^r\gamma_2 W_r^r(\nu, \rho)d\mu(\theta) \geq \gamma_2 t_r^r\left[\Delta^r + \frac{r}{4}\Delta^{r-1}\epsilon_n\right],\end{aligned}$$

  by Lemma 10. Furthermore, it is easy to show that

  $$\mathrm{SW}_r^r(P_{02}, Q_{02}) = \gamma_2\Delta^r \int_{\mathbb{S}^{d-1}} |\theta_1|^r d\mu(\theta) = \gamma_2(t_r\Delta)^r.$$

- **Construction 3.** We use the second part of Lemma 10. We set

  $$\nu = U(0, \bar{\beta}), \quad \rho = (1-\epsilon_m)U(0,1) + \epsilon_m\delta_1.$$

  Then, for all $\epsilon \in (0, 1]$,

  $$W_r(\nu, \rho) \geq \frac{1}{r+1}\left[|\Delta_{\bar{\beta}}|^r + r\epsilon_m|\Delta_{\bar{\beta}}|^{r-1}\right]$$

  We then obtain

  $$\begin{aligned}\mathrm{SW}_r^r(P_{13}, Q_{13}) &= \int_{\mathbb{S}^{d-1}} |\theta_1|^r s_2 W_r^r(\nu, \rho)d\mu(\theta) \\ &\geq \frac{t_r^r s_2}{r+1}\left[|\Delta_{\bar{\beta}}|^r + r\epsilon_m|\Delta_{\bar{\beta}}|^{r-1}\right].\end{aligned}$$

  On the other hand, it is easy to see that

  $$\mathrm{SW}_r^r(P_{03}, Q_{03}) = \frac{s_2 t_r^r|\Delta_{\bar{\beta}}|^r}{r+1}.$$

- **Construction 4.** We again use the second part of Lemma 10, setting

  $$\nu = U(0, \beta), \quad \rho = (1-\epsilon)U(0,1) + \epsilon_n\delta_1.$$

  The rest follows by the same argument as for Construction 3.

Computing the SJ_r evaluations. Our next step will be to compute the SJ_r functionals for the various distributions we have constructed. We note that for Construction 1 our distributions are allowed to have infinite SJ_r so we only need to consider Constructions 2-4. The calculations for Construction 3 and 4 follow along very similar lines to those of Construction 2, which we detail below.

- **Construction 2.** We have

$$\begin{aligned}\mathrm{SJ}_r(Q_{02}) = \mathrm{SJ}_r(Q_{12}) = \mathrm{SJ}_r(P_{02})&\\ &\leq \int_{\mathbb{S}^{d-1}} \int_0^1 \left(\frac{\sqrt{u(1-u)}}{1/(|\theta_1|\gamma_2^{1/r})} \right)^r du d\mu(\theta)\\ &\leq \int (|\theta_1|\gamma_2^{1/r})^r d\mu(\theta) \leq \gamma_2.\end{aligned}$$

  Furthermore,

$$\begin{aligned}\mathrm{SJ}_r(Q_{13}) &\leq \int_{\mathbb{S}^{d-1}} \int_0^1 \left(\frac{\sqrt{u(1-u)}}{(1-\epsilon_n)/(|\theta_1|\gamma_2^{1/r})} \right)^r du d\mu(\theta)\\ &\leq \frac{\gamma_2}{(1-\epsilon_n)^r} \int_0^1 [u(1-u)]^{\frac{r}{2}} du.\end{aligned}$$

  Choosing the constant $k_r > 0$ to satisfy $k_r < 1 - \left(\int_0^1 [u(1-u)]^{\frac{r}{2}} du\right)^{\frac{1}{r}}$ guarantees that the above display is bounded above by γ_2.

To complete the proof it remains to prove Lemma 10.

2.C.1.1 Proof of Lemma 10

We prove each of the two claims in turn.

Proof of Claim (1). Notice that the quantile functions of ν and ρ are respectively given by

$$\begin{aligned}&F^{-1}(u) = \begin{cases} \frac{u}{1+\epsilon}, & 0 \leq u \leq (1+\epsilon)/2, \\ \frac{1}{2} + \frac{u-(1+\epsilon)/2}{1-\epsilon}, & (1+\epsilon)/2 \leq u \leq 1 \end{cases},\\ &G^{-1}(u) = (u+\Delta)I(0 \leq u \leq 1).\end{aligned}$$

Thus,

$$\begin{aligned}W_r^r(\nu,\rho) &= \int_0^1 |F^{-1}(u) - G^{-1}(u)|^r du\\ &= \int_0^{\frac{1+\epsilon}{2}} \left| \Delta + u - \frac{u}{1+\epsilon} \right|^r du + \int_{\frac{1+\epsilon}{2}}^1 \left| \Delta + u - \frac{1}{2} - \frac{u-(1+\epsilon)/2}{1-\epsilon} \right|^r du\\ &= \text{(I)} + \text{(II)},\end{aligned}$$

say. We have,

$$(\text{I}) = \int_0^{(1+\epsilon)/2} \left[\Delta + \frac{\epsilon}{1+\epsilon} u\right]^r du = \frac{1+\epsilon}{\epsilon(r+1)} \left\{ \left(\Delta + \frac{\epsilon}{2}\right)^{r+1} - \Delta^{r+1} \right\}.$$

Also,

$$\begin{aligned}
(\text{II}) &= \int_{(1+\epsilon)/2}^1 \left(\Delta + \frac{\epsilon}{1-\epsilon} - u\frac{\epsilon}{1-\epsilon}\right)^r du \\
&= -\frac{1-\epsilon}{(r+1)\epsilon} \left\{ \left((\Delta + \frac{\epsilon}{1-\epsilon} - \frac{\epsilon}{1-\epsilon} \right)^{r+1} - \left(\Delta + \frac{\epsilon}{1-\epsilon} - \frac{(1+\epsilon)\epsilon}{2(1-\epsilon)}\right)^{r+1} \right\} \\
&= -\frac{1-\epsilon}{\epsilon(r+1)} \left\{ \Delta^{r+1} - \left(\Delta + \frac{\epsilon}{2}\right)^{r+1} \right\}.
\end{aligned}$$

Thus,

$$\begin{aligned}
(\text{I}) + (\text{II}) &= \left\{ \left(\Delta + \frac{\epsilon}{2}\right)^{r+1} - \Delta^{r+1} \right\} \left(\frac{1+\epsilon}{\epsilon(r+1)} + \frac{1-\epsilon}{\epsilon(r+1)} \right\} \\
&= \frac{2}{\epsilon(r+1)} \left\{ \left(\Delta + \frac{\epsilon}{2}\right)^{r+1} - \Delta^{r+1} \right\} \\
&= \frac{2}{\epsilon(r+1)} \left\{ \Delta^{r+1} + \frac{r+1}{2}\Delta^r \epsilon + \frac{r(r+1)}{8}(\Delta + \widetilde{\epsilon})^{r-1}\epsilon^2 - \Delta^{r+1} \right\},
\end{aligned}$$

for some $\widetilde{\epsilon} \in (0, \epsilon/2)$, by a first-order Taylor expansion. Therefore,

$$W_r^r(\nu, \rho) = \Delta^r + \frac{r}{4}(\Delta + \widetilde{\epsilon})^{r-1}\epsilon \geq \Delta^r + \frac{r}{4}\Delta^{r-1}\epsilon,$$

and the claim follows. □

Proof of Claim (2). The respective quantile functions of ν and ρ are given by

$$F^{-1}(u) = \begin{cases} \frac{u}{1-\epsilon}, & 0 \leq u \leq 1-\epsilon, \\ 1, & 1-\epsilon < u \leq 1 \end{cases}, \qquad G^{-1}(u) = \xi u I(0 \leq u \leq 1).$$

Thus,

$$\begin{aligned}
\text{SW}_r^r(\nu, \rho) &= \int \left|F^{-1}(u) - G^{-1}(u)\right|^r du \\
&= \int_0^{1-\epsilon} \left|\frac{u}{1-\epsilon} - \xi u\right|^r du + \int_{1-\epsilon}^1 |1 - \xi u|^r du \\
&= \int_0^{1-\epsilon} \left[\frac{u}{1-\epsilon} - \xi u\right]^r du + \int_{1-\epsilon}^1 [1 - \xi u]^r du, \quad (\text{Since } \xi \in (0,1]) \\
&= \frac{1-\epsilon}{r+1}(-\Delta_\xi + \epsilon\xi)^r + \frac{1}{\xi(r+1)}\left[(-\Delta_\xi + \epsilon\xi)^{r+1} - (-\Delta_\xi)^{r+1}\right]
\end{aligned}$$

$$
\begin{aligned}
&= \frac{1}{r+1}\left[(-\Delta_\xi + \epsilon\xi)^r \left(1 - \epsilon + \frac{-\Delta_\xi + \epsilon\xi}{\xi}\right) - \frac{|\Delta_\xi|^{r+1}}{\xi}\right] \\
&= \frac{1}{\xi(r+1)}\left[(|\Delta_\xi| + \epsilon\xi)^r - |\Delta_\xi|^{r+1}\right] \\
&= \frac{1}{\xi(r+1)}\left[(|\Delta_\xi| + \epsilon\xi)^r - |\Delta_\xi|^{r+1}\right] \\
&\geq \frac{1}{\xi(r+1)}\left[|\Delta_\xi|^r + r\epsilon\xi|\Delta_\xi|^{r-1} - |\Delta_\xi|^{r+1}\right] \\
&= \frac{|\Delta_\xi|^r}{r+1} + \frac{r\epsilon|\Delta_\xi|^{r-1}}{r+1}.
\end{aligned}
$$

The claim follows. □

2.D Proof of Theorem 6

By the same argument as in the proof of Proposition 4, it will suffice to prove equation (2.18). Let $P, Q \in \mathcal{K}_{r,\rho}(b)$. We prove the claim in five steps.

Step 0: Preparation. By the same argument as in the proof of Lemma 3, notice that there exists a constant $C_\rho > 0$ such that for any $\theta \in \mathbb{S}^{d-1}$,

$$
\left|F_\theta^{-1}(u)\right| \vee \left|G_\theta^{-1}(u)\right| \leq C_\rho \left(\frac{b_\theta}{u(1-u)}\right)^{\frac{1}{\rho}}, \quad u \in (0,1),
$$

where for $X \sim P$, $Y \sim Q$ and each $\theta \in \mathbb{S}^{d-1}$,

$$
b_\theta = \mathbb{E}[|X^\top \theta|^\rho] + \mathbb{E}[|Y^\top \theta|^\rho].
$$

Notice that the assumption $P, Q \in \mathcal{K}_{r,\rho}(b)$ for $\rho > 2r$ implies that $\int b_\theta^{\frac{r}{\rho}} d\mu(\theta) \leq 2b$. In particular, b_θ is finite for almost all $\theta \in \mathbb{S}^{d-1}$. These statements may be applied analogously to the empirical measures P_n and Q_m. Specifically, we have

$$
\left|F_{\theta,n}^{-1}(u)\right| \vee \left|G_{\theta,m}^{-1}(u)\right| \leq C_\rho \left(\frac{b_{\theta,nm}}{u(1-u)}\right)^{\frac{1}{\rho}}, \quad u \in (0,1),
$$

where we set

$$
b_{\theta,nm} := \int |x^\top \theta|^\rho dP_n(x) + \int |y^\top \theta|^\rho dQ_m(y).
$$

Combining these facts, we have for all $u \in (0,1)$,

$$
\left|F_\theta^{-1}(u)\right| \vee \left|G_\theta^{-1}(u)\right| \vee \left|F_{\theta,n}^{-1}(u)\right| \vee \left|G_{\theta,m}^{-1}(u)\right| \leq \psi_\theta(u) := C_\rho \left(\frac{b_\theta + b_{\theta,nm}}{u(1-u)}\right)^{\frac{1}{\rho}}.
$$

We suppress the dependence of ψ_θ on n and m for ease of notation, but we emphasize that $\psi_\theta(u)$ is a random variable.

Our proof makes use of a uniform self-normalized concentration inequality for the empirical quantile process, which was introduced in Section 2.B.2, as part of the proof of Lemma 5, and also in Example 2. Specifically, for any $\epsilon \in (0,1)$, let $\gamma_{\epsilon,n}, \eta_{\epsilon,n}$ be the sequences given in equation (2.42), with inverses given in equation (2.43), and defined in terms of the quantity

$$\nu_{\epsilon,n} = \sqrt{\frac{16}{n}\left[\log(16/\epsilon) + \log(2n+1)\right]},$$

for any given $\epsilon \in (0,1)$. Recall that for any $\theta \in \mathbb{S}^{d-1}$, the event

$$A_\epsilon = \Big\{ u \in (0,1) : F_\theta^{-1}(\gamma_{\epsilon,n}(u)) \le F_{\theta,n}^{-1}(u) \le F_\theta^{-1}(\eta_{\epsilon,n}(u)) \Big\}$$

satisfies $\mathbb{P}(A_\epsilon) \geq 1-\epsilon$.

Step 1: First Reduction. Apply a similar first-order Taylor expansion as in the proof of Lemma 6, to the cost function $|\cdot|^r$, to deduce that

$$\begin{aligned}
&\big|\mathrm{SW}_r^r(P_n, Q_m) - \mathrm{SW}_r^r(P,Q)\big| \\
&\qquad \le r \int_{\mathbb{S}^{d-1}} \int_0^1 |\widetilde{F}_{\theta,n}^{-1}(u) - \widetilde{G}_{\theta,m}^{-1}(u)|^{r-1} \\
&\qquad\qquad \times \Big[|F_{\theta,n}^{-1}(u) - F_\theta^{-1}(u)| + |G_{\theta,m}^{-1}(u) - G_\theta^{-1}(u)| \Big] du d\mu(\theta),
\end{aligned}$$

for some $\widetilde{F}_{\theta,n}^{-1}(u)$ on the segment joining $F_\theta^{-1}(u)$ and $F_{\theta,n}^{-1}(u)$, and some $\widetilde{G}_{\theta,m}^{-1}(u)$ on the segment joining $G_\theta^{-1}(u)$ and $G_{\theta,m}^{-1}(u)$. By definition of ψ_θ, we deduce

$$\begin{aligned}
&\big|\mathrm{SW}_r^r(P_n, Q_m) - \mathrm{SW}_r^r(P,Q)\big| \\
&\quad \lesssim \int_{\mathbb{S}^{d-1}} \int_0^1 \psi_\theta^{r-1}(u) \Big[|F_{\theta,n}^{-1}(u) - F_\theta^{-1}(u)| + |G_{\theta,m}^{-1}(u) - G_\theta^{-1}(u)| \Big] du d\mu(\theta) \\
&\quad = \int_{\mathbb{S}^{d-1}} \int_0^{1/2} \psi_\theta^{r-1}(u) \Big[|F_{\theta,n}^{-1}(u) - F_\theta^{-1}(u)| + |G_{\theta,m}^{-1}(u) - G_\theta^{-1}(u)| \Big] du d\mu(\theta) \\
&\quad + \int_{\mathbb{S}^{d-1}} \int_{1/2}^1 \psi_\theta^{r-1}(u) \Big[|F_{\theta,n}^{-1}(u) - F_\theta^{-1}(u)| + |G_{\theta,m}^{-1}(u) - G_\theta^{-1}(u)| \Big] du d\mu(\theta).
\end{aligned}$$

We shall bound the quantity

$$T = \int_{\mathbb{S}^{d-1}} \int_{1/2}^1 \psi_\theta^{r-1}(u) |F_{\theta,n}^{-1}(u) - F_\theta^{-1}(u)| du d\mu(\theta),$$

and a similar argument can be used to bound the remaining terms in the penultimate display.

Throughout the sequel, let $\epsilon \in (0,1)$ be chosen such that $\nu_{\epsilon,n} \le 1$. Let

$$\beta > \left(\frac{1}{2} - \frac{r}{\rho}\right)^{-1} > 0, \tag{2.46}$$

where we recall that $\rho > 2r$. Define $\delta_k = k^{-\beta}/2$ for all $k \geq 1$. Notice that $\delta_1 = 1/2$. Furthermore, let

$$K \equiv K_n = 1 \vee \left\lfloor (c\nu_{\epsilon,n})^{-\frac{\rho}{\beta(\rho-r)}} \right\rfloor$$

for a constant $c \geq 8$ to be specified below. We summarize a few algebraic facts in relation to the sequences $\delta_k, \eta_{\epsilon,n}, \gamma_{\epsilon,n}$, which we prove in Section 2.D.1.

Lemma 11. *There exists a choice of the constant $c \geq 8$, as well as a constant $c_1 > 0$, both possibly depending on β, ρ, r, such that for all $n \geq 1$, and all $\epsilon \in (0,1)$ for which $\nu_{\epsilon,n} \leq 1$, the following properties hold.*

(i) $c_1 \nu_{\epsilon,n}^{\frac{\rho}{\rho-r}} \geq \delta_K$. *Furthermore, if $K \geq 2$, then $\delta_K \geq (c\nu_{\epsilon,n})^{\frac{\rho}{\rho-r}}/2$.*

(ii) $1 - \gamma_{\epsilon,n}(1-\delta_k) \leq \delta_k + \frac{\nu_{\epsilon,n}^2}{1+\nu_{\epsilon,n}^2} + \nu_{\epsilon,n}\sqrt{\delta_k}$ *for all $k = 1, \dots, K$.*

(iii) $\frac{1-\delta_k}{1+\nu_{\epsilon,n}^2} \leq \eta_{\epsilon,n}(1-\delta_k) \leq 1 - \delta_k + \nu_{\epsilon,n}^2 + \nu_{\epsilon,n}\sqrt{\delta_k}$, *for all $k = 1, \dots, K$.*

(iv) $\eta_{\epsilon,n}(1-\delta_k) \leq \gamma_{\epsilon,n}(1-\delta_{k+1})$, *for all $k = 1, \dots, K-1$, if $K \geq 2$.*

With these facts in place, consider the decomposition

$$T = \int_{\mathbb{S}^{d-1}} \int_{1/2}^{1} \psi_\theta^{r-1}(u) |F_{\theta,n}^{-1}(u) - F_\theta^{-1}(u)| du d\mu(\theta) = \int_{\mathbb{S}^{d-1}} T_\theta d\mu(\theta), \tag{2.47}$$

where we recall that $\delta_1 = 1/2$, and we set

$$T_\theta = \int_{1/2}^{1} |F_{\theta,n}^{-1}(u) - F_\theta^{-1}(u)| \psi_\theta^{r-1}(u) du = T_{\theta,K}^* + \sum_{k=1}^{K-1} T_{\theta,k}, \quad \theta \in \mathbb{S}^{d-1}, \tag{2.48}$$

$$T_{\theta,k} = \int_{1-\delta_k}^{1-\delta_{k+1}} |F_{\theta,n}^{-1}(u) - F_\theta^{-1}(u)| \psi_\theta^{r-1}(u) du, \;\; k = 1, \dots, K-1, \tag{2.49}$$

$$T_{\theta,K}^* = \int_{1-\delta_K}^{1} |F_{\theta,n}^{-1}(u) - F_\theta^{-1}(u)| \psi_\theta^{r-1}(u) du. \tag{2.50}$$

In the following two steps, we bound the preceding two terms for any fixed $\theta \in \mathbb{S}^{d-1}$ and $k = 1, \dots, K-1$. The symbol "$\lesssim$" will always hide constants which do not depend on θ and k.

Step 2: Bounding $T_{\theta,K}^*$. We have,

$$\begin{aligned} T_{\theta,K}^* = \int_{1-\delta_K}^{1} |F_{\theta,n}^{-1}(u) - F_\theta^{-1}(u)| \psi_\theta^{r-1}(u) du &\lesssim \int_{1-\delta_K}^{1} \psi_\theta^r(u) du \\ &\lesssim (b_\theta + b_{\theta,nm})^{\frac{r}{\rho}} \int_{1-\delta_K}^{1} (1-u)^{-\frac{r}{\rho}} du \\ &\lesssim (b_\theta + b_{\theta,nm})^{\frac{r}{\rho}} \delta_K^{1-\frac{r}{\rho}} \lesssim (b_\theta + b_{\theta,nm})^{\frac{r}{\rho}} \nu_{\epsilon,n}, \end{aligned}$$

where the final inequality follows by Lemma 11(i).

Step 3: Bounding $T_{\theta,k}$ in Probability. The bulk of our work will now go into bounding $T_{\theta,k}$, for any given $\theta \in \mathbb{S}^{d-1}$ and $k = 1, \dots, K-1$. Notice that this case is vacuous when $K = 1$. The following derivations are performed over the event A_ϵ. By its definition, we have for any $k = 1, \dots, K-1$,

$$\begin{aligned}
T_{\theta,k} &\leq \int_{1-\delta_k}^{1-\delta_{k+1}} \left[F_\theta^{-1}(\eta_{\epsilon,n}(u)) - F_\theta^{-1}(\gamma_{\epsilon,n}(u))\right] \psi_\theta^{r-1}(u) du \\
&\leq \psi_\theta^{r-1}(1-\delta_{k+1}) \int_{1-\delta_k}^{1-\delta_{k+1}} \left[F_\theta^{-1}(\eta_{\epsilon,n}(u)) - F_\theta^{-1}(\gamma_{\epsilon,n}(u))\right] du \\
&= \psi_\theta^{r-1}(1-\delta_{k+1}) \Bigg[\int_{\eta_{\epsilon,n}(1-\delta_k)}^{\eta_{\epsilon,n}(1-\delta_{k+1})} F_\theta^{-1}(u) \frac{\partial \eta_{\epsilon,n}^{-1}(u)}{\partial u} du \\
&\qquad\qquad - \int_{\gamma_{\epsilon,n}(1-\delta_k)}^{\gamma_{\epsilon,n}(1-\delta_{k+1})} F_\theta^{-1}(u) \frac{\partial \gamma_{\epsilon,n}^{-1}(u)}{\partial u} du \Bigg] \\
&= \psi_\theta^{r-1}(1-\delta_{k+1})(A_{\theta,k} + B_{\theta,k}),
\end{aligned} \tag{2.51}$$

where

$$\begin{aligned}
A_{\theta,k} &= \int_{\eta_{\epsilon,n}(1-\delta_k)}^{\eta_{\epsilon,n}(1-\delta_{k+1})} F_\theta^{-1}(u) du - \int_{\gamma_{\epsilon,n}(1-\delta_k)}^{\gamma_{\epsilon,n}(1-\delta_{k+1})} F_\theta^{-1}(u) du, \\
B_{\theta,k} &= \int_{\eta_{\epsilon,n}(1-\delta_k)}^{\eta_{\epsilon,n}(1-\delta_{k+1})} F_\theta^{-1}(u) \left(\frac{\partial \eta_{\epsilon,n}^{-1}(u)}{\partial u} - 1 \right) du \\
&\quad - \int_{\gamma_{\epsilon,n}(1-\delta_k)}^{\gamma_{\epsilon,n}(1-\delta_{k+1})} F_\theta^{-1}(u) \left(\frac{\partial \gamma_{\epsilon,n}^{-1}(u)}{\partial u} - 1 \right) du.
\end{aligned}$$

Step 3.1: Bounding $A_{\theta,k}$. Consider the decomposition

$$A_{\theta,k} = \left(\int_{\eta_{\epsilon,n}(1-\delta_k)}^{\gamma_{\epsilon,n}(1-\delta_{k+1})} + \int_{\gamma_{\epsilon,n}(1-\delta_{k+1})}^{\eta_{\epsilon,n}(1-\delta_{k+1})} - \int_{\gamma_{\epsilon,n}(1-\delta_k)}^{\eta_{\epsilon,n}(1-\delta_k)} - \int_{\eta_{\epsilon,n}(1-\delta_k)}^{\gamma_{\epsilon,n}(1-\delta_{k+1})} \right) F_\theta^{-1}(u) du.$$

Using Lemma 11(iv) and the fact that $\eta_{\epsilon,n} \geq \gamma_{\epsilon,n}$, the four lower bounds of integration in the above display are less than their respective upper bounds. Therefore,

$$\begin{aligned}
A_{\theta,k} &= \left(\int_{\gamma_{\epsilon,n}(1-\delta_{k+1})}^{\eta_{\epsilon,n}(1-\delta_{k+1})} - \int_{\gamma_{\epsilon,n}(1-\delta_k)}^{\eta_{\epsilon,n}(1-\delta_k)} \right) F_\theta^{-1}(u) du \\
&\leq \left(\int_{\gamma_{\epsilon,n}(1-\delta_{k+1})}^{\eta_{\epsilon,n}(1-\delta_{k+1})} + \int_{\gamma_{\epsilon,n}(1-\delta_k)}^{\eta_{\epsilon,n}(1-\delta_k)} \right) \psi_\theta(u) du \\
&\leq \psi_\theta(\eta_{\epsilon,n}(1-\delta_{k+1})) \Big[\big(\eta_{\epsilon,n}(1-\delta_{k+1}) - \gamma_{\epsilon,n}(1-\delta_{k+1})\big) \\
&\qquad\qquad + \big(\eta_{\epsilon,n}(1-\delta_k) - \gamma_{\epsilon,n}(1-\delta_k)\big) \Big]
\end{aligned}$$

$$\begin{aligned}
&\leq \psi_\theta(\eta_{\epsilon,n}(1-\delta_{k+1}))\left[\frac{\nu_{\epsilon,n}\sqrt{\nu_{\epsilon,n}^2+4\delta_{k+1}}}{1+\nu_{\epsilon,n}^2}+\frac{\nu_{\epsilon,n}\sqrt{\nu_{\epsilon,n}^2+4\delta_k}}{1+\nu_{\epsilon,n}^2}\right]\\
&\lesssim \psi_\theta(\eta_{\epsilon,n}(1-\delta_{k+1}))\left[\nu_{\epsilon,n}^2+\nu_{\epsilon,n}\sqrt{\delta_k}\right].
\end{aligned}$$

By Lemma 11(i), we deduce that

$$A_{\theta,k}\lesssim \psi_\theta(\eta_{\epsilon,n}(1-\delta_{k+1}))\nu_{\epsilon,n}\left[\delta_K^{\frac{\rho-r}{\rho}}+\delta_k^{\frac{1}{2}}\right]\lesssim \psi_\theta(\eta_{\epsilon,n}(1-\delta_{k+1}))\nu_{\epsilon,n}\sqrt{\delta_k},$$

where we used the assumption $\rho > 2r$. We next turn to bounding $B_{\theta,k}$.

Step 3.2: Bounding $B_{\theta,k}$. We have,

$$\begin{aligned}
\frac{\partial}{\partial u}\eta_{\epsilon,n}^{-1}(u)&=1-\frac{\nu_{\epsilon,n}}{2}\left[\sqrt{\frac{1-u}{u}}-\sqrt{\frac{u}{1-u}}\right],\quad u\in[\eta_{\epsilon,n}(1/2),\eta_{\epsilon,n}(1)],\\
\frac{\partial}{\partial u}\gamma_{\epsilon,n}^{-1}(u)&=1+\frac{\nu_{\epsilon,n}}{2}\left[\sqrt{\frac{1-u}{u}}-\sqrt{\frac{u}{1-u}}\right],\quad u\in[\gamma_{\epsilon,n}(1/2),\gamma_{\epsilon,n}(1)].
\end{aligned}$$

Since $\nu_{\epsilon,n}\leq 1$, notice that $\gamma_{\epsilon,n}(1/2)$ and $\eta_{\epsilon,n}(1/2)$ are bounded below by a positive universal constant. Therefore, in the above display, the first terms in brackets are bounded above by positive universal constants, uniformly over the stated ranges, leading to

$$\begin{aligned}
\left|\frac{\partial}{\partial u}\eta_{\epsilon,n}^{-1}(u)-1\right|&\lesssim\frac{\nu_{\epsilon,n}}{2}\sqrt{\frac{1}{1-u}},\quad u\in[\eta_{\epsilon,n}(1/2),\eta_{\epsilon,n}(1)],\\
\left|\frac{\partial}{\partial u}\gamma_{\epsilon,n}^{-1}(u)-1\right|&\lesssim\frac{\nu_{\epsilon,n}}{2}\sqrt{\frac{1}{1-u}},\quad u\in[\gamma_{\epsilon,n}(1/2),\gamma_{\epsilon,n}(1)].
\end{aligned}$$

Deduce that,

$$\begin{aligned}
B_{\theta,k}&=\int_{\eta_{\epsilon,n}(1-\delta_k)}^{\eta_{\epsilon,n}(1-\delta_{k+1})}F_\theta^{-1}(u)\left(\frac{\partial\eta_{\epsilon,n}^{-1}(u)}{\partial u}-1\right)du\\
&\qquad-\int_{\gamma_{\epsilon,n}(1-\delta_k)}^{\gamma_{\epsilon,n}(1-\delta_{k+1})}F_\theta^{-1}(u)\left(\frac{\partial\gamma_{\epsilon,n}^{-1}(u)}{\partial u}-1\right)du\\
&\lesssim\nu_{\epsilon,n}\int_{\eta_{\epsilon,n}(1-\delta_k)}^{\eta_{\epsilon,n}(1-\delta_{k+1})}|F_\theta^{-1}(u)|\sqrt{\frac{1}{1-u}}\frac{du}{2}\\
&\qquad+\nu_{\epsilon,n}\int_{\gamma_{\epsilon,n}(1-\delta_k)}^{\gamma_{\epsilon,n}(1-\delta_{k+1})}|F_\theta^{-1}(u)|\sqrt{\frac{1}{1-u}}\frac{du}{2}\\
&\leq 2\nu_{\epsilon,n}\psi_\theta(\eta_{\epsilon,n}(1-\delta_{k+1}))\sqrt{1-\gamma_{\epsilon,n}(1-\delta_k)}\\
&\lesssim\nu_{\epsilon,n}\psi_\theta(\eta_{\epsilon,n}(1-\delta_{k+1}))\sqrt{\delta_k+\nu_{\epsilon,n}^2+\nu_{\epsilon,n}\sqrt{\delta_k}}\quad\text{(By Lemma 11(ii))}
\end{aligned}$$

$$\begin{aligned}
&\lesssim \nu_{\epsilon,n}\psi_\theta(\eta_{\epsilon,n}(1-\delta_{k+1}))\left[\sqrt{\delta_k}+\delta_k^{\frac{\rho-r}{\rho}}+\delta_k^{\frac{\rho-r}{2\rho}+\frac{1}{4}}\right] \quad \text{(By Lemma 11(i))}\\
&\lesssim \nu_{\epsilon,n}\psi_\theta(\eta_{\epsilon,n}(1-\delta_{k+1}))\sqrt{\delta_k},
\end{aligned}$$

where we again used the assumption $\rho > 2r$.

Step 4: Bounding T_θ in Probability. Combine the conclusions of Steps 3.1 and 3.2 with equation (2.51) to deduce

$$\begin{aligned}
&\sum_{k=1}^{K-1} T_{\theta,k}\\
&\lesssim \nu_{\epsilon,n}\sum_{k=1}^{K-1}\sqrt{\delta_k}\psi_\theta^{r-1}(1-\delta_{k+1})\psi_\theta(\eta_{\epsilon,n}(1-\delta_{k+1}))\\
&\lesssim \nu_{\epsilon,n}\sum_{k=1}^{K-1}\sqrt{\delta_k}\psi_\theta^{r}(\eta_{\epsilon,n}(1-\delta_{k+1}))\\
&\lesssim \nu_{\epsilon,n}(b_\theta+b_{\theta,nm})^{\frac{r}{\rho}}\sum_{k=1}^{K-1}\sqrt{\delta_k}\big[1-\eta_{\epsilon,n}(1-\delta_{k+1})\big]^{-\frac{r}{\rho}}\\
&\le \nu_{\epsilon,n}(b_\theta+b_{\theta,nm})^{\frac{r}{\rho}}\sum_{k=1}^{K-1}\sqrt{\delta_k}\left[\delta_k-\nu_{\epsilon_n}^2-\nu_{\epsilon,n}\sqrt{\delta_k}\right]^{-\frac{r}{\rho}} \quad \text{(By Lem. 11(iii))}\\
&\le \nu_{\epsilon,n}(b_\theta+b_{\theta,nm})^{\frac{r}{\rho}}\sum_{k=1}^{K-1}\sqrt{\delta_k}\left[\delta_k-\Big(\delta_K^{\frac{2(\rho-r)}{\rho}}+\delta_K^{\frac{\rho-r}{\rho}}\delta_k^{\frac{1}{2}}\Big)/4\right]^{-\frac{r}{\rho}} \quad \text{(By Lem. 11(i))}\\
&\lesssim \nu_{\epsilon,n}(b_\theta+b_{\theta,nm})^{\frac{r}{\rho}}\sum_{k=1}^{K-1}\delta_k^{\frac{1}{2}-\frac{r}{\rho}},
\end{aligned}$$

where we again used the fact that $\rho > 2r$ on the final line. By definition of β in equation (2.46), the sequence $\delta_k^{\frac{1}{2}-\frac{r}{\rho}}$ is summable, thus the summation in the final line of the above display is bounded above by a finite constant depending only on r, ρ, β. Combine this fact with the conclusion of Step 2 to deduce that, for a constant $C = C(\beta, \rho, r)$, we have

$$T_\theta \le C(b_\theta + b_{\theta,nm})^{\frac{r}{\rho}}\nu_{\epsilon,n}, \tag{2.52}$$

over the high-probability event A_ϵ, for all ϵ such that $\nu_{\epsilon,n} \le 1$.

Step 4: Bounding T_θ in Expectation. We now wish to turn the preceding display into a bound on the expectation $\mathbb{E}[T_\theta]$. Notice that $\mathbb{E}[b_{\theta,nm}] = b_\theta$, thus by Markov's inequality, we have for all $y > 0$,

$$\mathbb{P}(b_{\theta,nm}^{r/\rho} \ge b_\theta^{r/\rho}y) \le \mathbb{E}[b_{\theta,nm}]y^{-\rho/r}/b_\theta \le y^{-\rho/r}.$$

Furthermore, by inverting the definition of $\nu_{\epsilon,n}$ in terms of ϵ, equation (2.52) implies that for all $0 < u \leq 1$,

$$\mathbb{P}\left(T_\theta \geq Cu(b_\theta + b_{\theta,nm})^{r/\rho}\right) \leq \frac{2n+1}{16}\exp\left(-\frac{nu^2}{16}\right).$$

Combine the preceding two displays to deduce that for all $0 < u \leq 1$,

$$\begin{aligned}
&\mathbb{P}\left\{T_\theta > Cub_\theta^{r/\rho} + Cu(b_\theta/n)^{r/\rho}\exp\left(\frac{(r/\rho)nu^2}{16}\right)\right\} \\
&\quad= \mathbb{P}\left\{T_\theta > Cub_\theta^{\frac{r}{\rho}} + Cu\left(\frac{b_\theta}{n}\right)^{\frac{r}{\rho}} e^{\frac{(r/\rho)nu^2}{16}}, b_{\theta,nm}^{r/\rho} \leq (b_\theta/n)^{r/\rho}e^{\frac{(r/\rho)nu^2}{16}}\right\} \\
&\qquad+ \mathbb{P}\left\{T_\theta > Cub_\theta^{r/\rho} + Cu(b_\theta/n)^{r/\rho}e^{\frac{(r/\rho)nu^2}{16}}, b_{\theta,nm}^{r/\rho} > (b_\theta/n)^{r/\rho}e^{\frac{(r/\rho)nu^2}{16}}\right\} \\
&\quad\leq \mathbb{P}\left\{T_\theta > Cu(b_\theta^{\frac{r}{\rho}} + b_{\theta,nm}^{\frac{r}{\rho}})\right\} + \mathbb{P}\left\{b_{\theta,nm}^{r/\rho} > (b_\theta/n)^{r/\rho}\exp\left(\frac{(r/\rho)nu^2}{16}\right)\right\} \\
&\quad\leq \mathbb{P}\left\{T_\theta > Cu(b_\theta + b_{\theta,nm})^{\frac{r}{\rho}}\right\} + \mathbb{P}\left\{b_{\theta,nm}^{\frac{r}{\rho}} > (b_\theta/n)^{r/\rho}\exp\left(\frac{(r/\rho)nu^2}{16}\right)\right\} \\
&\quad\lesssim n\exp\left(-\frac{nu^2}{16}\right).
\end{aligned}$$

Let $f_n(u) = Cub_\theta^{r/\rho} + Cu(b_\theta/n)^{r/\rho}\exp\big((r/\rho)nu^2/16\big)$. f_n is strictly increasing over $\mathbb{R}_+$, thus it is invertible with inverse f_n^{-1}. Notice that $f_n^{-1}(0) = 0$ and $f_n^{-1}(u) \to \infty$ as $u \to \infty$. Furthermore,

$$f_n'(u) \lesssim b_\theta^{r/\rho} + (b_\theta/n)^{r/\rho}\Big[\exp\big((r/\rho)nu^2/16\big) + nu^2\exp\big((r/\rho)nu^2/16\big)\Big].$$

Now, let $t_n = \sqrt{\frac{16\log n}{n}}$ and let n be sufficiently large to ensure $t_n \leq 1$. We have,

$$\begin{aligned}
&\mathbb{E}\Big[T_\theta \cdot I(T_\theta \leq f_n(1))\Big] \\
&\qquad= \int_0^{f_n(1)} \mathbb{P}(T_\theta \geq x)dx \\
&\qquad\leq f_n(t_n) + \int_{f_n(t_n)}^{f_n(1)} \mathbb{P}(T_\theta \geq x)dx \\
&\qquad= f_n(t_n) + \int_{t_n}^{1} \mathbb{P}(T_\theta \geq f(u))f'(u)du \\
&\qquad\lesssim f_n(t_n) + b_\theta^{r/\rho}n\int_{t_n}^1 \exp\big(-nu^2/16\big)du \\
&\qquad\qquad+ \left(\frac{b_\theta}{n}\right)^{r/\rho} n\int_{t_n}^1 \exp\big(-n(1-r/\rho)u^2/16\big)du
\end{aligned}$$

$$
\begin{aligned}
&\quad + \left(\frac{b_\theta}{n}\right)^{r/\rho} n^2 \int_{t_n}^1 u^2 \exp\big(-n(1-r/\rho)u^2/16\big)du \\
&\lesssim f_n(t_n) + b_\theta^{r/\rho}\sqrt{n}\exp(-nt_n^2/16) \\
&\quad + \left(\frac{b_\theta}{n}\right)^{r/\rho} \sqrt{n}(1+\sqrt{n}t_n)\exp\big(-n(1-r/\rho)t_n^2/16\big),
\end{aligned}
$$

where the bound on the final term can be obtained by integration by parts. Thus,

$$
\begin{aligned}
&\mathbb{E}\Big[T_\theta \cdot I(T_\theta \leq f_n(1))\Big] \\
&\qquad \lesssim b_\theta^{\frac{r}{\rho}}\sqrt{\frac{\log n}{n}} + \frac{b_\theta^{r/\rho}}{\sqrt{n}} + \left(\frac{b_\theta}{n}\right)^{r/\rho}\sqrt{n\log n}\cdot n^{\frac{r}{\rho}-1} \lesssim b_\theta^{\frac{r}{\rho}}\sqrt{\frac{\log n}{n}}.
\end{aligned}
$$

Finally, in order to control $\mathbb{E}\left[T_\theta \cdot I(T_\theta > f_n(1))\right]$, we use the naive bound

$$
\begin{aligned}
T_\theta &= \int_{1/2}^1 |F_{\theta,n}^{-1}(u) - F_\theta^{-1}(u)|\psi_\theta^{r-1}(u)du \\
&\lesssim \int_{1/2}^1 \psi_\theta^r(u)du \lesssim (b_{\theta,nm}+b_\theta)^{\frac{r}{\rho}} \int_{1/2}^1 \frac{du}{(1-u)^{r/\rho}} \lesssim (b_{\theta,nm}+b_\theta)^{\frac{r}{\rho}}.
\end{aligned}
$$

Using the inequality $2r < \rho$ and Jensen's inequality, we deduce

$$
\mathbb{E}\left[T_\theta^2\right] \lesssim \mathbb{E}[(b_\theta + b_{\theta,nm})^{2r/\rho}] = b_\theta^{\frac{2r}{\rho}} + \mathbb{E}[b_{\theta,nm}]^{\frac{2r}{\rho}} = 2b_\theta^{\frac{2r}{\rho}}.
$$

Thus, using the Cauchy-Schwarz inequality, we arrive at

$$
\mathbb{E}\Big[T_\theta \cdot I(T_\theta > f_n(1))\Big] \leq \sqrt{\mathbb{E}\left[T_\theta^2\right]\mathbb{P}\big(T_\theta > f_n(1)\big)} \lesssim b_\theta^{\frac{r}{\rho}}\sqrt{n}\exp(-n/32).
$$

We deduce that

$$
\mathbb{E}[T_\theta] = \mathbb{E}\Big[T_\theta \cdot I(T_\theta > f_n(1))\Big] + \mathbb{E}\Big[T_\theta \cdot I(T_\theta \leq f_n(1))\Big] \lesssim b_\theta^{\frac{r}{\rho}}(\log n/n)^{1/2}.
$$

Step 5: Conclusion. By the Fubini-Tonelli Theorem and the nonnegativity of T_θ, we deduce from the above display that,

$$
\begin{aligned}
\mathbb{E}[T] &= \mathbb{E}\left[\int_{\mathbb{S}^{d-1}} T_\theta d\mu(\theta)\right] \\
&= \int_{\mathbb{S}^{d-1}} \mathbb{E}\left[T_\theta\right] d\mu(\theta) \lesssim (\log n/n)^{1/2} \int_{\mathbb{S}^{d-1}} b_\theta^{r/\rho} d\mu(\theta) \leq b(\log n/n)^{1/2},
\end{aligned}
$$

where the final inequality follows from the assumption that $P \in \mathcal{K}_{r,\rho}(b)$. The claim follows. □

2.D.1 Proof of Lemma 11

Part (i) is trivial from the definition of K. To prove parts (ii)–(iv), note that for all $k \geq 1$,

$$\begin{aligned}\gamma_{\epsilon,n}(1-\delta_k) &= \frac{2(1-\delta_k) + \nu_{\epsilon,n}^2 - \nu_{\epsilon,n}\sqrt{\nu_{\epsilon,n}^2 + 4\delta_k(1-\delta_k)}}{2(1+\nu_{\epsilon,n}^2)} \\ &\geq \frac{1-\delta_k}{1+\nu_{\epsilon,n}^2} - \nu_{\epsilon,n}\sqrt{\delta_k}. \end{aligned} \tag{2.53}$$

Therefore,

$$\gamma_{\epsilon,n}(1-\delta_k) \geq 1 - \delta_k - \frac{\nu_{\epsilon,n}^2}{1+\nu_{\epsilon,n}^2} - \nu_{\epsilon,n}\sqrt{\delta_k},$$

which proves claim (ii). Furthermore,

$$\begin{aligned}\frac{1-\delta_k}{1+\nu_{\epsilon,n}^2} \leq \eta_{\epsilon,n}(1-\delta_k) &= \frac{2(1-\delta_k) + \nu_{\epsilon,n}^2 + \nu_{\epsilon,n}\sqrt{\nu_{\epsilon,n}^2 + 4\delta_k(1-\delta_k)}}{2(1+\nu_{\epsilon,n}^2)} \\ &\leq (1-\delta_k) + \frac{1}{2}\left[2\nu_{\epsilon,n}^2 + \nu_{\epsilon,n}\sqrt{4\delta_k}\right] \\ &= (1-\delta_k) + \nu_{\epsilon,n}^2 + \nu_{\epsilon,n}\sqrt{\delta_k}, \end{aligned} \tag{2.54}$$

which proves claim (iii). To prove claim (iv), it follows from equations (2.53)–(2.54) that it suffices to show

$$\frac{1-\delta_{k+1}}{1+\nu_{\epsilon,n}^2} - \nu_{\epsilon,n}\sqrt{\delta_{k+1}} \geq (1-\delta_k) + \nu_{\epsilon_n}^2 + \nu_{\epsilon,n}\sqrt{\delta_k}.$$

This assertion is equivalent to

$$\delta_k - \delta_{k+1} \geq \nu_{\epsilon,n}\left[\sqrt{\delta_{k+1}} + \sqrt{\delta_k}\right] + \nu_{\epsilon,n}^2\left[2 - \delta_k + \nu_{\epsilon_n}^2 + \nu_{\epsilon,n}\sqrt{\delta_k} + \nu_{\epsilon,n}\sqrt{\delta_{k+1}}\right],$$

which, in turn, will be satisfied if the following inequality holds,

$$\delta_k - \delta_{k+1} \geq 2\nu_{\epsilon,n}\sqrt{\delta_k} + 5\nu_{\epsilon,n}^2.$$

Using a first-order Taylor expansion of the map $x \mapsto x^{-\beta}$, notice that $\delta_k - \delta_{k+1} \geq \beta(k+1)^{-(1+\beta)}/2$. This fact together with property (i) implies that it is enough to show

$$\frac{\beta}{2}(k+1)^{-(1+\beta)} \geq \frac{2(2\delta_K)^{\frac{\rho-r}{\rho}}}{c}\frac{k^{-\frac{\beta}{2}}}{\sqrt{2}} + \frac{5}{c^2}(2\delta_K)^{2\frac{\rho-r}{\rho}},$$

for which, in turn, it suffices to show

$$\frac{\beta}{2}(k+1)^{-(1+\beta)} \geq \frac{2k^{-\beta\left(\frac{3}{2}-\frac{r}{\rho}\right)}}{c\sqrt{2}} + \frac{5}{c^2}k^{-2\beta\frac{\rho-r}{\rho}}.$$

By definition of β in condition (2.46), we have

$$1 + \beta < \beta\left(\frac{3}{2} - \frac{r}{\rho}\right) \vee 2\beta\frac{\rho - r}{\rho},$$

thus for a sufficiently large choice of c, depending only on β, ρ, r, the penultimate display holds for all $k \geq 1$. The claim follows. □

2.E Proof of Theorem 7

We begin by formally stating assumptions **B1**-**B3**, referenced in the statement of Theorem 7.

B1 $\gamma_{\epsilon,n}(u), \eta_{\epsilon,n}(u)$, viewed as functions of $u \in [0,1]$, are nondecreasing, differentiable, invertible with differentiable inverses, and are also respectively nondecreasing and nonincreasing functions of $\epsilon \in (0,1)$ and $n \geq 1$.

B2 There exists a constant $K_1 > 0$ such that for all $f, g \in \{\gamma_{\tau/N,n}, \eta_{\tau/N,n}, \iota : \tau \in \{\epsilon, \epsilon\wedge\alpha\}\}$, with ι the identity function on $[0,1]$, we have $\delta/2 \leq f(\delta)$, $f(1-\delta) \leq 1 - \delta/2$, and

$$\sup_{\delta/2 \leq u \leq 1-\delta/2} \left| \frac{\partial g^{-1}(f(u))}{\partial u} - 1 \right| \leq K_1 \kappa_{\varepsilon,n}.$$

B3 There exists a constant $K_2 > 1$ such that for all $t \in [\delta/2, 1-\delta/2]$ and all $\gamma_{\varepsilon,n}(t) \leq u \leq \eta_{\varepsilon,n}(t)$ we have

$$\frac{1}{K_2} \leq \frac{\gamma_{\varepsilon,n}^{-1}(u) - \eta_{\varepsilon,n}^{-1}(u)}{\eta_{\varepsilon,n}(t) - \gamma_{\varepsilon,n}(t)} \leq K_2.$$

Proof of Theorem 7. Throughout the proof, the symbol "$\lesssim$" is used to hide constants possibly depending on K_1, K_2, δ_0, r. Furthermore, the symbol $\varkappa_N$ is used to denote a random variable depending only on $\theta_1, \dots, \theta_N$, whose definition may change from line to line, but which always satisfies $\mathbb{E}_{\mu^{\otimes N}}[\varkappa_N] \leq CN^{-1/2r} I(d \geq 2)$, where $C > 0$ denotes a constant possibly depending on K_1, K_2, δ, b, r. We prove the claim in five steps.

Step 0: Setup. With probability at least $1 - \epsilon$, uniformly in $j = 1, \dots, N$ and $u \in (0,1)$, we have both

$$F_{\theta_j,n}^{-1}\big(\gamma_{\epsilon/N,n}(u)\big) \leq F_{\theta_j}^{-1}(u) \leq F_{\theta_j,n}^{-1}\big(\eta_{\epsilon/N,n}(u)\big), \tag{2.55}$$

and,

$$G_{\theta_j,m}^{-1}\big(\gamma_{\epsilon/N,m}(u)\big) \leq G_{\theta_j}^{-1}(u) \leq G_{\theta_j,m}^{-1}\big(\eta_{\epsilon/N,m}(u)\big). \tag{2.56}$$

All derivations which follow will be carried out on the event that the above two inequalities are satisfied, which has probability at least $1 - \epsilon$. For notational simplicty, we will write $a = \alpha/N$, $e = \epsilon/N$, and we recall that $\varepsilon = e \wedge a$.

Recall that for all $\theta \in \{\theta_1, \dots, \theta_N\}$, $F_{\theta,n}, G_{\theta,m}$ denote the empirical CDFs of P_θ and Q_θ respectively, and $F_{\theta,n}^{-1}, G_{\theta,m}^{-1}$ their corresponding quantile functions. We may write

$$C_{nm}^{(N)} = \Bigg[\left(\frac{1}{1-2\delta} \int_{\mathbb{S}^{d-1}} \int_\delta^{1-\delta} A_{\theta,nm}^r(u) du d\mu_N(\theta) \right)^{\frac{1}{r}}, \left(\frac{1}{1-2\delta} \int_{\mathbb{S}^{d-1}} \int_\delta^{1-\delta} B_{\theta,nm}^r(u) du d\mu_N(\theta) \right)^{\frac{1}{r}} \Bigg],$$

where

$$A_{\theta,nm}(u) = \big[F_{\theta,n}^{-1}\big(\gamma_{a,n}(u)\big) - G_{\theta,m}^{-1}\big(\eta_{a,m}(u)\big)\big] \vee \big[G_{\theta,m}^{-1}\big(\gamma_{a,m}(u)\big) - F_{\theta,n}^{-1}\big(\eta_{a,n}(u)\big)\big] \vee 0,$$

and

$$B_{\theta,nm}(u)=\big[F_{\theta,n}^{-1}\big(\eta_{a,n}(u)\big) - G_{\theta,m}^{-1}\big(\gamma_{a,m}(u)\big)\big]\vee\big[G_{\theta,m}^{-1}\big(\eta_{a,m}(u)\big) - F_{\theta,n}^{-1}\big(\gamma_{a,n}(u)\big)\big].$$

We will first show that

$$\left| \frac{1}{1-2\delta} \int_{\mathbb{S}^{d-1}} \int_\delta^{1-\delta} A_{\theta,nm}^r(u) du d\mu_N(\theta) - \left[\mathrm{SW}_{r,\delta}^{(N)}(P,Q)\right]^r \right| \lesssim \psi_{\varepsilon,nm} + \varphi_{\varepsilon,nm} + \varkappa_N.$$

A similar argument can be used to bound this expression with $A_{\theta,nm}$ replaced by $B_{\theta,nm}$, and will lead to the claim.

We will assume without loss of generality that $r > 1$ in what follows. As will be clear from the proof, the arguments of Steps 2 and 4 alone may easily be used to prove the claim when $r = 1$.

Step 1: Taylor Expansion Setup. Let $u \in [\delta, 1-\delta]$, and $\theta \in \{\theta_1, \dots, \theta_N\}$. By Taylor expansion of the map $(x, y) \in \mathbb{R}^2 \mapsto (x-y)^r$, there exists $\widetilde{F}_{\theta,n}^{-1}(u)$ (resp. $\widetilde{G}_{\theta,m}^{-1}(u)$) on the segment joining $F_{\theta,n}^{-1}(\gamma_{a,n}(u))$ and $F_\theta^{-1}(u)$ (resp. $G_\theta^{-1}(u)$ and $G_{\theta,m}^{-1}(\eta_{a,m}(u))$) such that

$$\left[F_{\theta,n}^{-1}(\gamma_{a,n}(u)) - G_{\theta,m}^{-1}(\eta_{a,m}(u))\right]^r = \left[F_\theta^{-1}(u) - G_\theta^{-1}(u)\right]^r + \xi_{nm}(u),$$

where

$$\xi_{\theta,nm}(u) = r\left(\widetilde{F}_{\theta,n}^{-1}(u) - \widetilde{G}_{\theta,m}^{-1}(u)\right)^{r-1} \times \Big\{ \Big(F_{\theta,n}^{-1}\big(\gamma_{a,n}(u)\big) - F_\theta^{-1}(u)\Big) - \Big(G_{\theta,m}^{-1}\big(\eta_{a,m}(u)\big) - G_\theta^{-1}(u)\Big) \Big\}.$$

Likewise, there exists $\overline{F}_{\theta,n}^{-1}(u)$ (resp. $\overline{G}_{\theta,m}^{-1}(u)$) on the segment joining $F_\theta^{-1}(u)$ and $F_{\theta,n}^{-1}\big(\eta_{a,n}(u)\big)$ (resp. $G_{\theta,m}^{-1}\big(\gamma_{a,m}(u)\big)$ and $G_\theta^{-1}(u)$) such that

$$\left[G_{\theta,m}^{-1}(\gamma_{a,m}(u)) - F_{\theta,n}^{-1}\big(\eta_{a,n}(u)\big)\right]^r = \left[G_\theta^{-1}(u) - F_\theta^{-1}(u)\right]^r + \zeta_{\theta,nm}(u),$$

where

$$\zeta_{\theta,nm}(u) = r\Big(\overline{G}^{-1}_{\theta,m}(u) - \overline{F}^{-1}_{\theta,n}(u)\Big)^{r-1} \times \\ \Big\{ \Big(F^{-1}_{\theta,n}\big(\eta_{a,n}(u)\big) - F^{-1}_{\theta}(u)\Big) - \Big(G^{-1}_{\theta,m}\big(\gamma_{a,m}(u)\big) - G^{-1}_{\theta}(u)\Big) \Big\}.$$

Now, consider the numerical inequality $\big|(a^r + b) \vee ((-a)^r + d) \vee 0 - |a|^r\big| \leq 3(|b| + |d|)$, for all $a \in \mathbb{R}$, $b, d \in \mathbb{R}$, Taking $a = (F^{-1}_\theta - G^{-1}_\theta)$, $b = \xi_{\theta,nm}$ and $d = \zeta_{\theta,nm}$, this inequality implies

$$\begin{aligned}
&\left| \frac{1}{1-2\delta} \int_{\mathbb{S}^{d-1}} \int_\delta^{1-\delta} A^r_{\theta,nm}(u) du d\mu_N(\theta) - \Big[\mathrm{SW}^{(N)}_{r,\delta}(P,Q)\Big]^r \right| \\
&\leq \frac{1}{1-2\delta} \int_{\mathbb{S}^{d-1}} \int_\delta^{1-\delta} \Bigg| \Big[\big(F^{-1}_\theta(u) - G^{-1}_\theta(u)\big)^r + \xi_{\theta,nm}(u)\Big] \\
&\vee \Big[\big(G^{-1}_\theta(u) - F^{-1}_\theta(u)\big)^r + \zeta_{\theta,nm}(u)\Big] \vee 0 - \big|F^{-1}_\theta(u) - G^{-1}_\theta(u)\big|^r \Bigg| du d\mu_N(\theta) \\
&\lesssim \int_{\mathbb{S}^{d-1}} \int_\delta^{1-\delta} |\zeta_{\theta,nm}(u)| du d\mu_N(\theta) + \int_{\mathbb{S}^{d-1}} \int_\delta^{1-\delta} |\xi_{\theta,nm}(u)| du d\mu_N(\theta). \qquad (2.57)
\end{aligned}$$

It will now suffice to bound the second term of the above display, and a similar bound will hold for the first. Note that

$$\int_{\mathbb{S}^{d-1}} \int_\delta^{1-\delta} |\xi_{\theta,nm}(u)| du d\mu_N(\theta) \leq r(\mathcal{I} + \mathcal{J}),$$

where,

$$\mathcal{I} = \int_{\mathbb{S}^{d-1}} \int_\delta^{1-\delta} \big|\widetilde{F}^{-1}_{\theta,n}(u) - \widetilde{G}^{-1}_{\theta,m}(u)\big|^{r-1} \big|F^{-1}_{\theta,n}\big(\gamma_{a,n}(u)\big) - F^{-1}_\theta(u)\big| du d\mu_N(\theta) \qquad (2.58)$$

$$\mathcal{J} = \int_{\mathbb{S}^{d-1}} \int_\delta^{1-\delta} \big|\widetilde{F}^{-1}_{\theta,n}(u) - \widetilde{G}^{-1}_{\theta,m}(u)\big|^{r-1} \big|G^{-1}_{\theta,m}\big(\eta_{a,m}(u)\big) - G^{-1}_\theta(u)\big| du d\mu_N(\theta). \qquad (2.59)$$

It will suffice to prove that $\mathcal{I} \lesssim \psi_{\varepsilon,nm}$ and $\mathcal{J} \lesssim \varphi_{\varepsilon,nm}$, up to terms depending only on N. We consider the cases $\mathrm{SJ}_{r,\delta/2}(P) \vee \mathrm{SJ}_{r,\delta/2}(Q) = \infty$ and $\mathrm{SJ}_{r,\delta/2}(P) \vee \mathrm{SJ}_{r,\delta/2}(Q) < \infty$ seperately.

Step 2: Bounding $\mathcal{I}$ and $\mathcal{J}$ when $\mathbf{SJ}_{r,\delta/2}(P) \vee \mathbf{SJ}_{r,\delta/2}(Q) = \infty$. We have,

$$\mathcal{I} \lesssim \int_{\mathbb{S}^{d-1}} \left(\sup_{\delta \leq u \leq 1-\delta} \left| \widetilde{F}^{-1}_{\theta,n}(u) - \widetilde{G}^{-1}_{\theta,m}(u) \right|^{r-1} \right) \times \\ \left(\int_\delta^{1-\delta} \left| F^{-1}_{\theta,n}(\gamma_{a,n}(u)) - F^{-1}_\theta(u) \right| du \right) d\mu_N(\theta).$$

We will bound each factor in the above integral, beginning with the second. Using inequality (2.55), since $e \geq \varepsilon$, we have for all $u \in [\delta, 1-\delta]$ and $\theta \in \{\theta_1, \dots, \theta_N\}$,

$$|F^{-1}_{\theta,n}(\gamma_{a,n}(u)) - F^{-1}_\theta(u)| \\ \leq \big[F^{-1}_\theta\big(\gamma^{-1}_{\varepsilon,n}(\gamma_{a,n}(u))\big) - F^{-1}_\theta(u)\big] + \big[F^{-1}_\theta(u) - F^{-1}_\theta\big(\eta^{-1}_{\varepsilon,n}(\gamma_{a,n}(u))\big)\big].$$

Now, write $x_n = \gamma_{\varepsilon,n}^{-1}(\gamma_{a,n}(\delta))$ and $y_n = \eta_{\varepsilon,n}^{-1}(\gamma_{a,n}(1-\delta))$. By condition **B2** and the definition of $\kappa_{\varepsilon,n}$, we have

$$|x_n - \delta| \vee |y_n - 1 - \delta| \leq \kappa_{\varepsilon,n}, \tag{2.60}$$

which, combined with the assumption that $\kappa_{\varepsilon,n} \leq \frac{\delta}{2} \wedge (1-2\delta)$, also implies

$$\delta \leq x_n \leq 1 - \delta \leq y_n \leq 1 - \delta/2.$$

Thus, for all $\theta \in \{\theta_1, \dots, \theta_N\}$,

$$\begin{aligned}
&\int_\delta^{1-\delta} \left[F_\theta^{-1}\big(\gamma_{\varepsilon,n}^{-1}(\gamma_{a,n}(u))\big) - F_\theta^{-1}(u)\right] du \\
&= \int_{x_n}^{y_n} F_\theta^{-1}(u) \left(\frac{\partial \gamma_{a,n}^{-1}(\gamma_{\varepsilon,n}(u))}{\partial u}\right) du - \int_\delta^{1-\delta} F_\theta^{-1}(u) du \\
&\leq \int_{x_n}^{y_n} F_\theta^{-1}(u) du + K_1 \kappa_{\varepsilon,n} \int_{x_n}^{y_n} |F_\theta^{-1}(u)| du - \int_\delta^{1-\delta} F_\theta^{-1}(u) du, \qquad \text{(By } \mathbf{B2}\text{)} \\
&\leq \int_{1-\delta}^{y_n} F_\theta^{-1}(u) du - \int_\delta^{x_n} F_\theta^{-1}(u) du + K_1 \kappa_{\varepsilon,n} |F_\theta^{-1}(y_n)| \\
&\lesssim (y_n - 1 + \delta)\Big[|F_\theta^{-1}(y_n)| \vee |F_\theta^{-1}(1-\delta)|\Big] \\
&+ (x_n - \delta)\Big[|F_\theta^{-1}(\delta)| \vee |F_\theta^{-1}(x_n)|\Big] + \kappa_{\varepsilon,n} |F_\theta^{-1}(y_n)| \\
&\lesssim (y_n - 1 + \delta + \kappa_{\varepsilon,n}) |F_\theta^{-1}(1-\delta/2)| + (x_n - \delta)|F_\theta^{-1}(\delta)|.
\end{aligned}$$

Since $P \in \overline{\mathcal{K}}_2(b)$, the quantiles in the above display are bounded above by a universal multiple of $(b/\delta)^{1/2}$, by Lemma 3. Thus, together with equation (2.60), we arrive at

$$\int_\delta^{1-\delta} \left[F_\theta^{-1}\big(\gamma_{\varepsilon,n}^{-1}(\gamma_{a,n}(u))\big) - F_\theta^{-1}(u)\right] du \lesssim \kappa_{\varepsilon,n} (b/\delta)^{1/2},$$

We similarly have,

$$\int_\delta^{1-\delta} \left[F_\theta^{-1}(u) - F_\theta^{-1}\big(\eta_{\varepsilon,n}^{-1}(\gamma_{a,n}(u))\big)\right] du \lesssim \kappa_{\varepsilon,n} (b/\delta)^{1/2}.$$

Combining these facts, we obtain

$$\mathcal{I} \lesssim \kappa_{\varepsilon,n} (b/\delta)^{1/2} \int_{\mathbb{S}^{d-1}} \sup_{\delta \leq u \leq 1-\delta} \left|\widetilde{F}_{\theta,n}^{-1}(u) - \widetilde{G}_{\theta,m}^{-1}(u)\right|^{r-1} d\mu_N(\theta). \tag{2.61}$$

We now bound the second factor in the above display. Since we have $\widetilde{F}_{\theta,n}^{-1}(u) \in [F_{\theta,n}^{-1}(\gamma_{a,n}(u)), F_\theta^{-1}(u)]$ and $\widetilde{G}_{\theta,m}^{-1}(u) \in [G_\theta^{-1}(u), G_{\theta,m}^{-1}(\eta_{a,m}(u))]$, we deduce that for any $u \in [\delta, 1-\delta]$,

$$\begin{aligned}
&\left|\widetilde{F}_{\theta,n}^{-1}(u) - \widetilde{G}_{\theta,m}^{-1}(u)\right| \\
&\leq \left|G_\theta^{-1}(u) - F_\theta^{-1}(u)\right| + \left|F_\theta^{-1}(u) - \widetilde{F}_{\theta,n}^{-1}(u)\right| + \left|\widetilde{G}_{\theta,m}^{-1}(u) - G_\theta^{-1}(u)\right|
\end{aligned} \tag{2.62}$$

$$\begin{aligned}&\le \Big|G_\theta^{-1}(u) - F_\theta^{-1}(u)\Big| \\ &\quad + \Big|F_\theta^{-1}(u) - F_\theta^{-1}\big(\eta_{\varepsilon,n}^{-1}(\gamma_{a,n}(u))\big)\Big| + \Big|G_\theta^{-1}\big(\gamma_{\varepsilon,m}^{-1}(\eta_{a,m}(u))\big) - G_\theta^{-1}(u)\Big|.\end{aligned} \tag{2.63}$$

It follows that

$$\begin{aligned}&\int_{\mathbb{S}^{d-1}} \sup_{\delta\le u\le 1-\delta} \Big|\widetilde{F}_{\theta,n}^{-1}(u) - \widetilde{G}_{\theta,m}^{-1}(u)\Big|^{r-1} d\mu_N(\theta) \\ &\qquad\qquad \lesssim \Big\{\mathrm{SW}_{\infty,\delta}^{(r-1)}(P,Q) + U_{\varepsilon,n}(P) + U_{\varepsilon,m}(Q)\Big\} + \varkappa_N.\end{aligned}$$

We conclude this section of the proof by combining the above display with equation (2.61). We then have

$$\mathcal{I} \lesssim \kappa_{\varepsilon,n}(b/\delta)^{1/2}\Big\{\mathrm{SW}_{\infty,\delta}^{(r-1)}(P,Q) + U_{\varepsilon,n}(P) + U_{\varepsilon,m}(Q) + \varkappa_N\Big\} \lesssim \psi_{\varepsilon,nm} + \varkappa_N,$$

and by a symmetric argument,

$$\mathcal{J} \lesssim \varphi_{\varepsilon,nm} + \varkappa_N.$$

Step 3: Bounding $\mathcal{I}$ and $\mathcal{J}$ when $\mathrm{SJ}_{r,\delta/2}(P)\vee\mathrm{SJ}_{r,\delta/2}(Q) < \infty$. By means of Hölder's inequality, we have

$$\begin{aligned}\mathcal{I} &\le \int_{\mathbb{S}^{d-1}}\int_\delta^{1-\delta} \big|\widetilde{F}_{\theta,n}^{-1}(u) - \widetilde{G}_{\theta,m}^{-1}(u)\big|^{r-1}\big|F_{\theta,n}^{-1}(\gamma_{a,n}(u)) - F_\theta^{-1}(u)\big| du d\mu_N(\theta) \\ &\le \left(\int_{\mathbb{S}^{d-1}}\int_\delta^{1-\delta} \big|\widetilde{F}_{\theta,n}^{-1}(u) - \widetilde{G}_{\theta,m}^{-1}(u)\big|^r du d\mu_N(\theta)\right)^{\frac{r-1}{r}} \times \\ &\qquad \left(\int_{\mathbb{S}^{d-1}}\int_\delta^{1-\delta} \big|F_{\theta,n}^{-1}\big(\gamma_{a,n}(u)\big) - F_\theta^{-1}(u)\big|^r du d\mu_N(\theta)\right)^{\frac{1}{r}} \\ &\lesssim \left(\int_{\mathbb{S}^{d-1}}\int_\delta^{1-\delta} \big|\widetilde{F}_{\theta,n}^{-1}(u) - \widetilde{G}_{\theta,m}^{-1}(u)\big|^r du d\mu_N(\theta)\right)^{\frac{r-1}{r}} \times \\ &\qquad \left(\int_{\mathbb{S}^{d-1}}\int_\delta^{1-\delta} \big[F_\theta^{-1}\big(\gamma_{\varepsilon,n}^{-1}(\gamma_{a,n}(u))\big) - F_\theta^{-1}\big(\eta_{\varepsilon,n}^{-1}(\gamma_{a,n}(u))\big)\big]^r du d\mu_N(\theta)\right)^{\frac{1}{r}} \\ &= \left(\int_{\mathbb{S}^{d-1}}\int_\delta^{1-\delta} \big|\widetilde{F}_{\theta,n}^{-1}(u) - \widetilde{G}_{\theta,m}^{-1}(u)\big|^r du d\mu_N(\theta)\right)^{\frac{r-1}{r}} \times \\ &\left(\int_{\mathbb{S}^{d-1}}\int_{\gamma_{a,n}(\delta)}^{\gamma_{a,n}(1-\delta)} \big[F_\theta^{-1}\big(\gamma_{\varepsilon,n}^{-1}(u)\big) - F_\theta^{-1}(\eta_{\varepsilon,n}^{-1}(u))\big]^r \left(\frac{\partial\gamma_{a,n}^{-1}(u)}{\partial u}\right) du d\mu_N(\theta)\right)^{\frac{1}{r}} \\ &\le \left(\int_{\mathbb{S}^{d-1}}\int_\delta^{1-\delta} \big|\widetilde{F}_{\theta,n}^{-1}(u) - \widetilde{G}_{\theta,m}^{-1}(u)\big|^r du d\mu_N(\theta)\right)^{\frac{r-1}{r}} \times \\ &\left(\int_{\mathbb{S}^{d-1}}\int_{\gamma_{a,n}(\delta)}^{\gamma_{a,n}(1-\delta)} \big[F_\theta^{-1}\big(\gamma_{\varepsilon,n}^{-1}(u)\big) - F_\theta^{-1}(\eta_{\varepsilon,n}^{-1}(u))\big]^r (1+K_1\kappa_{\varepsilon,n}) du d\mu_N(\theta)\right)^{\frac{1}{r}},\end{aligned}$$

where we used condition **B2** on the final line. We now reason similarly as in Section 2.B.2. Since $\mathrm{SJ}_{r,\delta/2}(P) < \infty$, it follows from Lemma 2 that F_θ^{-1} is absolutely continuous over $[\delta/2, 1-\delta/2]$ for μ-almost every $\theta \in \mathbb{S}^{d-1}$. We thus have for have almost surely that for each $\theta \in \{\theta_1, \dots, \theta_N\}$,

$$\begin{aligned}
&\int_{\gamma_{a,n}(\delta)}^{\gamma_{a,n}(1-\delta)} \left[F_\theta^{-1}\big(\gamma_{\varepsilon,n}^{-1}(u)\big) - F_\theta^{-1}(\eta_{\varepsilon,n}^{-1}(u))\right]^r du \\
&\quad = \int_{\gamma_{a,n}(\delta)}^{\gamma_{a,n}(1-\delta)} \left(\int_{\eta_{\varepsilon,n}^{-1}(u)}^{\gamma_{\varepsilon,n}^{-1}(u)} \frac{dt}{p_\theta(F_\theta^{-1}(t))}\right)^r du \\
&\quad \leq \int_{\gamma_{a,n}(\delta)}^{\gamma_{a,n}(1-\delta)} (\gamma_{\varepsilon,n}^{-1}(u) - \eta_{\varepsilon,n}^{-1}(u))^{r-1} \int_{\eta_{\varepsilon,n}^{-1}(u)}^{\gamma_{\varepsilon,n}^{-1}(u)} \left(\frac{1}{p_\theta(F_\theta^{-1}(t))}\right)^r dt du \\
&\quad \leq \int_{\delta/2}^{1-\delta/2} \left(\int_{\gamma_{\varepsilon,n}(t)}^{\eta_{\varepsilon,n}(t)} (\gamma_{\varepsilon,n}^{-1}(u) - \eta_{\varepsilon,n}^{-1}(u))^{r-1} du\right) \left(\frac{1}{p_\theta(F_\theta^{-1}(t))}\right)^r dt \\
&\quad \leq \int_{\delta/2}^{1-\delta/2} \left(\sup_{\gamma_{\varepsilon,n}(t) \leq u \leq \eta_{\varepsilon,n}(t)} (\gamma_{\varepsilon,n}^{-1}(u) - \eta_{\varepsilon,n}^{-1}(u))^{r-1}\right) \frac{\eta_{\varepsilon,n}(t) - \gamma_{\varepsilon,n}(t)}{[p_\theta(F_\theta^{-1}(t))]^r} dt \\
&\quad \lesssim \int_{\delta/2}^{1-\delta/2} \left(\frac{\eta_{\varepsilon,n}(t) - \gamma_{\varepsilon,n}(t)}{p_\theta(F_\theta^{-1}(t))}\right)^r dt \lesssim \int_{\delta/2}^{1-\delta/2} \left(\frac{\widetilde{\kappa}_{\varepsilon,n}(t)}{p_\theta(F_\theta^{-1}(t))}\right)^r dt,
\end{aligned}$$

where we repeatedly used assumption **B3** on the final line. We thus arrive at,

$$\mathcal{I} \lesssim \left(\int_{\mathbb{S}^{d-1}} \int_\delta^{1-\delta} \left|\widetilde{F}_{\theta,n}^{-1}(u) - \widetilde{G}_{\theta,m}^{-1}(u)\right|^r du d\mu_N(\theta)\right)^{\frac{r-1}{r}} \left[V_{\varepsilon,n}(P)\right]^{\frac{1}{r}} + \varkappa_N. \tag{2.64}$$

By using similar calculations as in equations (2.62) and (2.64) to bound the first factor in the above display, we have

$$\begin{aligned}
&\int_{\mathbb{S}^{d-1}} \int_\delta^{1-\delta} \left(\widetilde{F}_{\theta,n}^{-1}(u) - \widetilde{G}_{\theta,m}^{-1}(u)\right)^r du d\mu_N(\theta) \\
&\qquad \lesssim \mathrm{SW}_{r,\delta}^r(P,Q) + \int_{\mathbb{S}^{d-1}} \int_\delta^{1-\delta} \left|F_\theta^{-1}(u) - F_\theta^{-1}\big(\eta_{\varepsilon,n}^{-1}(\gamma_{a,n}(u))\big)\right|^r du d\mu_N(\theta) \\
&\qquad\qquad + \int_{\mathbb{S}^{d-1}} \int_\delta^{1-\delta} \left|G_\theta^{-1}\big(\gamma_{\varepsilon,m}^{-1}(\gamma_{a,m}(u))\big) - G_\theta^{-1}(u)\right|^r du d\mu_N(\theta) \\
&\qquad \lesssim \mathrm{SW}_{r,\delta}^r(P,Q) + V_{\varepsilon,n}(P) + V_{\varepsilon,m}(Q) + \varkappa_N,
\end{aligned}$$

Putting these facts together with equation (2.64), we arrive at

$$\mathcal{I} \lesssim \left(\mathrm{SW}_{r,\delta}^r(P,Q) + V_{\varepsilon,n}(P) + V_{\varepsilon,m}(Q)\right)^{\frac{r-1}{r}} \left[V_{\varepsilon,n}(P)\right]^{\frac{1}{r}} + \varkappa_N = \psi_{\varepsilon,nm} + \varkappa_N.$$

Finally, by a symmetric argument, we also have $\mathcal{J} \lesssim \varphi_{\varepsilon,nm} + \varkappa_N$.

Step 4: Conclusion. Returning to equation (2.57), we have shown

$$\int_{\mathbb{S}^{d-1}} \int_{\delta}^{1-\delta} |\xi_{nm}(u)| du d\mu_N(\theta) \lesssim \psi_{\varepsilon,nm} + \varphi_{\varepsilon,nm} + \varkappa_N.$$

By the same arguments, we may obtain the same upper bound, up to universal constant factors, on the integral $\int_{\mathbb{S}^{d-1}} \int_{\delta}^{1-\delta} |\zeta_{nm}(u)| du d\mu_N(\theta)$ in equation (2.57). We deduce that, for some $c_1 > 0$ (possibly depending on r), we have

$$\begin{aligned}\left(\frac{1}{1-2\delta} \int_{\mathbb{S}^{d-1}} \int_{\delta}^{1-\delta} A_{nm}^r(u) du d\mu_N(\theta)\right)^{\frac{1}{r}} \\ &\geq \left\{\left[\mathrm{SW}_{r,\delta}^{(N)}(P,Q)\right]^r - c_1\big(\psi_{\varepsilon,nm} + \varphi_{\varepsilon,nm} + \varkappa_N\big)\right\}^{\frac{1}{r}} \\ &\geq \left\{\mathrm{SW}_{r,\delta}^r(P,Q) - c_1\big(\psi_{\varepsilon,nm} + \varphi_{\varepsilon,nm} + \varkappa_N\big)\right\}^{\frac{1}{r}}.\end{aligned}$$

By the same arguments, there exists a constant $c_2 > 0$ such that

$$\begin{aligned}\left(\frac{1}{1-2\delta} \int_{\mathbb{S}^{d-1}} \int_{\delta}^{1-\delta} B_{nm}^r(u) du d\mu_N(\theta)\right)^{\frac{1}{r}} \\ &\leq \left\{\mathrm{SW}_{r,\delta}^r(P,Q) + c_2\big(\psi_{\varepsilon,nm} + \varphi_{\varepsilon,nm} + \varkappa_N\big)\right\}^{\frac{1}{r}}.\end{aligned}$$

The claim follows by choosing $c = c_1 \vee c_2$. □

2.F Proof of Theorem 8

The proof of this result has two main components. In Lemma 13 we show that the Sliced Wasserstein distance is Hadamard differentiable under certain conditions. Theorem 8 then follows via an application of the functional delta method.

Hadamard Differentiability of the Sliced Wasserstein Distance. Throughout this subsection, for a metric space (T,ρ), $C[T]$ denotes the set of real-valued continuous functions defined on T, endowed with the supremum norm, and $\ell^\infty(T) = \{f : T \to \mathbb{R} : \sup_{t\in T}|f(t)| < \infty\}$. Let $D[I]$ denote the space of càdlàg functions defined over an interval $I = [a_1, a_2] \subseteq \mathbb{R}$, endowed with the supremum norm. We will make use of the following result from van der Vaart and Wellner (1996) (Lemma 3.9.20), regarding the Hadamard differentiability of the quantile function at a fixed point $u \in (a_1, a_2)$. Let $\mathbb{D}_\psi$ denote the set of nondecreasing maps $A \in D[I]$ such that the set $\{x \in I : A(x) \geq u\}$ is nonempty for any given $u \in (0,1)$, and define the map

$$\psi : \mathbb{D}_\psi \subseteq D[I] \to \mathbb{R}, \quad \psi : A \mapsto A^{-1}(u) = \inf\{x \in I : A(x) \geq u\}. \tag{2.65}$$

Lemma 12 (van der Vaart and Wellner (1996)). *Let $A \in \mathbb{D}_\psi$ satisfy the following two properties.*

(i) A is differentiable at a point $\xi_u \in (a_1, a_2)$ such that $A(\xi_u) = u$.

(ii) A has strictly positive derivative at ξ_u.

Then, ψ is Hadamard-differentiable at A tangentially to the set of functions $H \in D[I]$ which are continuous at ξ_u, with Hadamard derivative given by

$$\psi'_A(H) = -\frac{H(\xi_u)}{A'(\xi_u)}.$$

Now, define $\mathcal{H} = \mathbb{R} \times \mathbb{S}^{d-1}$, identified with the set of half-spaces in $\mathbb{R}^d$. Let $\mathbb{D} = \ell^\infty(\mathcal{H})$, and let $\mathbb{D}_0$ denote the subspace of $\mathbb{D}$ consisting of maps $F : \mathcal{H} \to \mathbb{R}$ such that $F(\cdot, \theta) \in C[\mathbb{R}]$ for all $\theta \in \mathbb{S}^{d-1}$. Furthermore, define $\mathbb{D}_\phi$ as the subset of maps $F : \mathcal{H} \to \mathbb{R}$ such that $F(\cdot, \theta) \in D[\mathbb{R}]$ is a CDF for all $\theta \in \mathbb{S}^{d-1}$. Define the map

$$\phi : \mathbb{D}_\phi^2 \to \mathbb{R}_+, \quad \phi : (F, G) \mapsto \frac{1}{1-2\delta} \int_{\mathbb{S}^{d-1}} \int_\delta^{1-\delta} \left|F^{-1}(u,\theta) - G^{-1}(u,\theta)\right|^r du d\mu(\theta),$$

for a fixed constant $\delta \in [0, 1/2)$, where we interchangeably employ the notation $F^{-1}(u,\theta) = F_\theta^{-1}(u) = \inf\{x \in \mathbb{R} : F_\theta(x) \geq u\}$ and $F(\cdot,\theta) = F_\theta(\cdot)$ in this section only.

The Hadamard differentiability of ϕ, tangentially to $\mathbb{D}_0$, is established below.

Lemma 13. *Assume the same conditions as in Theorem 8. Then, the map ϕ is Hadamard differentiable at (F, G), tangentially to $\mathbb{D}_0$, with Hadamard derivative given by*

$$\begin{aligned}
&\phi' : \mathbb{D}_0^2 \to \mathbb{R}, \\
&\phi'(H_1, H_2) = \frac{r}{1-2\delta} \int_{\mathbb{S}^{d-1}} \int_\delta^{1-\delta} \operatorname{sgn}\left(F^{-1}(u,\theta) - G^{-1}(u,\theta)\right) \times \\
&\left|F^{-1}(u,\theta) - G^{-1}(u,\theta)\right|^{r-1} \left(\frac{H_2(G^{-1}(u,\theta),\theta)}{q_\theta(G^{-1}(u,\theta))} - \frac{H_1(F^{-1}(u,\theta),\theta)}{p_\theta(F^{-1}(u,\theta))}\right) du d\mu(\theta).
\end{aligned}$$

Proof of Lemma 13. Let $(H_{1k})_{k=1}^\infty, (H_{2k})_{k=1}^\infty \subseteq \mathbb{D}$ be sequences satisfying $F + t_k H_{1k}, G + t_k H_{2k} \in \mathbb{D}_\phi$ for all $k \geq 1$, and such that H_{jk} converges uniformly to $H_j \in \mathbb{D}_0$, $j = 1, 2$. Let $t_k \downarrow 0$ as $k \to \infty$, and define for all $k \geq 1$,

$$\begin{aligned}
\Delta_k = \frac{1}{t_k(1-2\delta)} \int_{\mathbb{S}^{d-1}} \int_\delta^{1-\delta} \Big\{&\left|(F + t_k H_{1k})^{-1}(u,\theta) - (G + t_k H_{2k})^{-1}(u,\theta)\right|^r \\
&- \left|F^{-1}(u,\theta) - G^{-1}(u,\theta)\right|^r\Big\} du d\mu(\theta).
\end{aligned}$$

We will prove that the limit of Δ_k exists when taking $k \to \infty$. For all $r > 1$, the map $(x, y) \in \mathbb{R}^2 \mapsto |x - y|^r$ is continuously differentiable. Thus, for all $u \in [\delta, 1-\delta]$ and all $\theta \in \mathbb{S}^{d-1}$, there exists $\widetilde{F}_k^{-1}(u,\theta)$ (resp, $\widetilde{G}_k^{-1}(u,\theta)$) on the line joining $F^{-1}(u,\theta)$ (resp. $G^{-1}(u,\theta)$) and $(F + t_k H_{1k})^{-1}(u,\theta)$ (resp. $(G + t_k H_{2k})^{-1}(u,\theta)$) such that

$$\begin{aligned}
\Delta_k = \frac{r}{t_k(1-2\delta)} \int_{\mathbb{S}^{d-1}} \int_\delta^{1-\delta} &\varphi\big(\widetilde{F}_k^{-1}(u,\theta); \widetilde{G}_k^{-1}(u,\theta)\big) \\
&\times \Big\{\left[(F + t_k H_{1k})^{-1}(u,\theta) - F^{-1}(u,\theta)\right] \\
&- \left[(G + t_k H_{2k})^{-1}(u,\theta) - G^{-1}(u,\theta)\right]\Big\} du d\mu(\theta),
\end{aligned} \tag{2.66}$$

where $\varphi(x;y) = \mathrm{sgn}(x-y)|x-y|^{r-1}$. We will now argue that each of the limits

$$\begin{aligned}
B_1(u,\theta) &= \lim_{k\to\infty} B_{1k}(u,\theta), \quad B_{1k}(u,\theta) = \varphi\big(\widetilde{F}_k^{-1}(u,\theta); \widetilde{G}_k^{-1}(u,\theta)\big),\\
B_2(u,\theta) &= \lim_{k\to\infty} B_{2k}(u,\theta), \quad B_{2k}(u,\theta) = \frac{(F+t_kH_{1k})^{-1}(u,\theta) - F^{-1}(u,\theta)}{t_k},\\
B_3(u,\theta) &= \lim_{k\to\infty} B_{3k}(u,\theta), \quad B_{3k}(u,\theta) = \frac{(G+t_kH_{2k})^{-1}(u,\theta) - G^{-1}(u,\theta)}{t_k},
\end{aligned}$$

exist for $(\lambda\otimes\mu)$-almost every $(u,\theta)\in[\delta,1-\delta]\times\mathbb{S}^{d-1}$. Throughout the sequel, we write $I_\theta = [F^{-1}(\delta/2,\theta), F^{-1}(1-\delta/2,\theta)]$ for all $\theta\in\mathbb{S}^{d-1}$. By Lemma 3, there exist finite constants $a_1 < a_2$, depending on δ and the finite second moments of P and Q, such that $\bigcup_{\theta\in\mathbb{S}^{d-1}} I_\theta \subseteq [a_1,a_2]$. We fix $I=[a_1,a_2]$ in what follows.

Regarding the limit B_1, we make use of the following observation, which we prove below in Section 2.F.1. The conclusion of this assertion is stronger than necessary, but will be needed again in the sequel.

Lemma 14. *Assume the same conditions as Theorem 8. Then, for all $u\in[\delta,1-\delta]$, we have as $k\to\infty$,*

$$\begin{aligned}
\sup_{\theta\in\mathbb{S}^{d-1}} |(F+t_kH_{1k})^{-1}(u,\theta) - F^{-1}(u,\theta)| \to 0,\\
\sup_{\theta\in\mathbb{S}^{d-1}} |(G+t_kH_{2k})^{-1}(u,\theta) - G^{-1}(u,\theta)| \to 0.
\end{aligned}$$

Notice further that the map φ is continuous in both of its arguments, therefore we obtain from Lemma 14 that

$$B_1(u,\theta) = \varphi\big(F^{-1}(u,\theta); G^{-1}(u,\theta)\big),$$

for $(\lambda\otimes\mu)$-almost every (u,θ).

We now turn to the limit B_2. Recall that $A := F(\cdot,\theta)$ is absolutely continuous for any given $\theta\in\mathbb{S}^{d-1}$. For any fixed $u\in[\delta,1-\delta]$, the existence of $B_2(u,\theta)$ would be implied by the Hadamard differentiability of the map ψ in equation (2.65) at A, tangentially to $\mathbb{D}_0$, sufficient conditions for which are given by conditions (i) and (ii) of Lemma 12. Condition (i) is immediately satisfied for almost all $u\in[\delta,1-\delta]$ due to the absolute continuity of A. Furthermore, the assumption $J_{\infty,\delta/2}(P)<\infty$ implies that p_θ is nonzero at $A^{-1}(u)$ for almost every $u\in[\delta,1-\delta]$, implying that condition (ii) of Lemma 12 is satisfied for all such u. We deduce from Lemma 12 the limit

$$B_2(u,\theta) = -\frac{H_1(A^{-1}(u),\theta)}{p_\theta(A^{-1}(u))} = -\frac{H_1(F^{-1}(u,\theta),\theta)}{p_\theta(F^{-1}(u,\theta))},$$

for $(\lambda\otimes\mu)$-almost every (u,θ). We similarly obtain, almost everywhere,

$$B_3(u,\theta) = -\frac{H_2(G^{-1}(u,\theta),\theta)}{q_\theta(G^{-1}(u,\theta))}.$$

With these facts in place, the claim will follow from the Dominated Convergence Theorem if we are able to interchange the limit as $k \to \infty$ with the integrations in equation (2.66). To this end, it will suffice to show that there exists $K \geq 1$ such that

$$\operatorname*{esssup}_{\theta \in \mathbb{S}^{d-1}} \operatorname*{esssup}_{\delta \leq u \leq 1-\delta} \sup_{k \geq K} |B_{2k}(u, \theta)| < \infty. \tag{2.67}$$

A similar argument may then be used to obtain the same conclusion with B_{2k} replaced by B_{3k}. These facts will imply, in particular, that the maps $(F + t_k H_{1k})^{-1}$ and $(G + t_k H_{2k})^{-1}$ are uniformly bounded over $[\delta, 1-\delta] \times \mathbb{S}^{d-1}$, which, together with the continuity of φ, then also implies that the above display holds with B_{2k} replaced by B_{1k}.

It thus remains to prove equation (2.67). We shall make use of the following properties.

Lemma 15. *Under the assumptions of Theorem 8, the following assertions hold.*

(i) The family $\{F(\cdot, \theta)\}_{\theta \in \mathbb{S}^{d-1}}$ is uniformly absolutely continuous over I, in the sense that for all $t > 0$, there exists $\epsilon(t) > 0$ such that for any $[\alpha, \beta] \subseteq I$, we have

$$(\beta - \alpha) \leq \epsilon(t) \implies \sup_{\theta \in \mathbb{S}^{d-1}} \big| F(\alpha, \theta) - F(\beta, \theta) \big| \leq t.$$

(ii) We have,

$$\gamma := \inf_{\theta \in \mathbb{S}^{d-1}} \inf_{\delta/2 \leq u \leq 1 - 3\delta/4} \Big[F^{-1}(u + \delta/4, \theta) - F^{-1}(u, \theta) \Big] > 0.$$

(iii) There exists a constant $K \geq 1$ such that for all $k \geq K$, $u \in [\delta, 1-\delta]$, and $\theta \in \mathbb{S}^{d-1}$,

$$F^{-1}(3\delta/4, \theta) \leq (F + t_k H_{1k})^{-1}(u, \theta) \leq F^{-1}(1 - \delta/4, \theta).$$

(iv) In particular, $(F + t_k H_{1k})^{-1}(u, \theta) - \gamma \in I_\theta$ for all $\theta \in \mathbb{S}^{d-1}$.

Now, for all $u \in [\delta, 1-\delta]$ and $\theta \in \mathbb{S}^{d-1}$, write

$$\xi_{\theta,u} = F^{-1}(u, \theta), \quad \xi_{\theta,uk} = (F + t_k H_{1k})^{-1}(u, \theta). \tag{2.68}$$

For all $k \geq 1$, let $\epsilon_k := \epsilon(t_k)$ be the constant corresponding to the choice $t = t_k$ in the statement of Lemma 15(i). This defines a sequence $(\epsilon_k)_{k \geq 1}$, which we may assume is nonincreasing, and satisfies $\epsilon_k < \gamma$ for all $k \geq K$, without loss of generality. By definition of the quantile function, we have

$$(F + t_k H_{1k})(\xi_{\theta,uk} - \epsilon_k, \theta) \leq u \leq (F + t_k H_{1k})(\xi_{\theta,uk}, \theta).$$

We use these inequalities to bound $\xi_{\theta,uk} - \xi_{\theta,u}$. Assume first that $\xi_{\theta,uk} < \xi_{\theta,u}$. Then, by absolute continuity of $F(\cdot, \theta)$ for all $\theta \in \mathbb{S}^{d-1}$, we have

$$0 \leq (F + t_k H_{1k})(\xi_{\theta,uk}, \theta) - F(\xi_{\theta,u}, \theta) = t_k H_{1k}(\xi_{\theta,uk}, \theta) - \int_{\xi_{\theta,uk}}^{\xi_{\theta,u}} p_\theta(x) dx.$$

By Lemma 15(iii), we have $[\xi_{\theta,uk}, \xi_{\theta,u}] \subseteq I_\theta$ for all $k \geq K$, $u \in [\delta, 1-\delta]$, and $\theta \in \mathbb{S}^{d-1}$. Therefore, we obtain

$$0 \leq t_k H_{1k}(\xi_{\theta,uk}, \theta) - (\xi_{\theta,u} - \xi_{\theta,uk}) J^{-1}_{\infty,\delta/2}(P_\theta).$$

On the other hand, if $\xi_{\theta,uk} \geq \xi_{\theta,u}$, we have

$$\begin{aligned}
0 &\geq (F + t_k H_{1k})(\xi_{\theta,uk} - \epsilon_k) - F(\xi_{\theta,u}) \\
&= \Big[F(\xi_{\theta,uk} - \epsilon_k) - F(\xi_{\theta,uk})\Big] + \Big[F(\xi_{\theta,uk}) - F(\xi_{\theta,u})\Big] + t_k H_{1k}(\xi_{\theta,uk} - \epsilon_k) \\
&= -\int_{\xi_{\theta,uk}-\epsilon_k}^{\xi_{\theta,uk}} p_\theta(x)dx + \int_{\xi_{\theta,u}}^{\xi_{\theta,uk}} p_\theta(x)dx + t_k H_{1k}(\xi_{\theta,uk} - \epsilon_k) \\
&\geq -t_k + (\xi_{\theta,uk} - \xi_{\theta,u}) J^{-1}_{\infty,\delta/2}(P_\theta) + t_k H_{1k}(\xi_{\theta,uk} - \epsilon_k),
\end{aligned}$$

where on the final line, we lower bounded the first term as follows. Since $\epsilon_k < \gamma$, we have $\xi_{\theta,uk} - \epsilon_k \in I_\theta \subseteq I$ by Lemma 15(iv). Again, we also have $\xi_{\theta,uk} \in I$ by Lemma 15(iii). Therefore, by the definition of $\epsilon_k = \epsilon(t_k)$, Lemma 15(i) can be applied to obtain the stated lower bound. Combine the preceding two displays to deduce,

$$\begin{aligned}
|B_{2k}(u,\theta)| = \left|\frac{\xi_{\theta,uk} - \xi_{\theta,u}}{t_k}\right| &\leq (1 + |H_{1k}(\xi_{\theta,uk})| + |H_{1k}(\xi_{\theta,uk} - \epsilon_k)|)\, J_{\infty,\delta/2}(P_\theta) \\
&\leq (1 + \|H_{1k}\|_\infty + \|H_{2k}\|_\infty) \sup_{\theta \in \mathbb{S}^{d-1}} J_{\infty,\delta/2}(P_\theta).
\end{aligned}$$

Now, recall that for $j = 1, 2$, H_{jk} converges uniformly to $H_j \in \mathbb{D}_0 \subseteq \ell^\infty(\mathcal{H})$. It must also follow that, up to modifying the value of $K \geq 1$, the function H_{jk} is bounded, uniformly in $k \geq K$. This fact combined with our assumption on P implies that the right-hand side of the above display is bounded above by a finite constant not depending on k, u, θ. Equation (2.67) readily follows, leading to the claim. □

Proof of Theorem 8. The claim consists of two statements, to be proven in parallel. Since the set of half-spaces $\mathcal{H}$ forms a separable Vapnik-Chervonenkis class, it is Donsker, implying that the empirical process $\mathbb{G}_{nm} = \sqrt{\frac{nm}{n+m}}(P_n - P, Q_m - Q)$ converges weakly in $\mathbb{D}^2 = \ell^\infty(\mathcal{H}) \times \ell^\infty(\mathcal{H})$,

$$\sup_{h \in \mathrm{BL}_1(\mathbb{D}^2)} \Big|\mathbb{E}[h(\mathbb{G}_{nm})] - \mathbb{E}[h(\mathbb{G}_{(P,Q)})]\Big| \longrightarrow 0, \tag{2.69}$$

to a process $\mathbb{G}_{(P,Q)} := (\sqrt{a}\mathbb{G}_P, \sqrt{1-a}\mathbb{G}_Q)$, where $\mathbb{G}_P$ and $\mathbb{G}_Q$ denote independent P- and Q-Brownian bridges respectively, and where we identify the set $\mathcal{H}$ with the set of indicator functions over $\mathcal{H}$. Under this abuse of notation, notice that the process $\mathbb{G}_P$ takes the form $\mathbb{G}_P(x,\theta) = \mathbb{G} \circ F(x,\theta)$ for a standard Brownian Bridge $\mathbb{G}$, for all $(x,\theta) \in \mathcal{H}$. By assumption, for all $\theta \in \mathbb{S}^{d-1}$, $F(\cdot,\theta)$ is continuous, and since almost all sample paths of $\mathbb{G}$ are continuous, we deduce that almost every sample path of $\mathbb{G}_P(\cdot,\theta)$ is also continuous. We deduce that

$\mathbb{G}_P$ takes values in $\mathbb{D}_0$ almost surely, and similarly for $\mathbb{G}_Q = \overline{\mathbb{G}} \circ G$, where $\overline{\mathbb{G}}$ is a standard Brownian Bridge independent of $\mathbb{G}$.

Furthermore, Theorem 3.6.3 of van der Vaart and Wellner (1996) implies the same conditional limiting distribution for the bootstrap empirical process $\mathbb{G}^*_{nm} = \sqrt{\frac{nm}{n+m}}(P^*_n - P_n, Q^*_m - Q_m)$,

$$\begin{gathered}\sup_{h\in \mathrm{BL}_1(\mathbb{D}^2)} \Big|\mathbb{E}\big[h(\mathbb{G}^*_{nm})\big|X_1,\dots,X_n,Y_1,\dots,Y_m\big] - \mathbb{E}\big[h(\mathbb{G}_{(P,Q)})\big]\Big| \to 0,\\ \mathbb{E}\big[h(\mathbb{G}^*_{nm})|X_1,\dots,X_n,Y_1,\dots,Y_m\big]^* - \mathbb{E}\big[h(\mathbb{G}^*_{nm})|X_1,\dots,X_n,Y_1,\dots,Y_m\big]_* \to 0,\end{gathered} \tag{2.70}$$

in outer probability, where h ranges over $\mathrm{BL}_1(\mathbb{D}^2)$. Now, the Hadamard differentiability of ϕ (Lemma 13) together with equation (2.69) and the functional delta method (see for instance Theorems 3.9.4 of van der Vaart and Wellner (1996)) implies

$$\sup_{h\in \mathrm{BL}_1(\mathbb{R})} \left|\mathbb{E}\left[h\left(\sqrt{\frac{nm}{n+m}}(\phi(P_n,Q_m) - \phi(P,Q))\right)\right] - \mathbb{E}\Big[h\left(\phi'\left(\mathbb{G}_{(P,Q)}\right)\right)\Big]\right| \to 0. \tag{2.71}$$

Likewise, the delta method for the bootstrap (Theorem 3.9.11 of van der Vaart and Wellner (1996)) and equations (2.69) and (2.70) imply

$$\begin{aligned}\sup_{h\in \mathrm{BL}_1(\mathbb{R})} \Bigg|\mathbb{E}\left[h\left(\sqrt{\frac{nm}{n+m}}\left(\phi(P^*_n,Q^*_m) - \phi(P_n,Q_m)\right)\right)\Bigg|X_1,\dots,X_n,Y_1,\dots,Y_m\right]\\ - \mathbb{E}\Big[h\left(\phi'\left(\mathbb{G}_{(P,Q)}\right)\right)\Big]\Bigg| \longrightarrow 0,\end{aligned} \tag{2.72}$$

in outer probability. A combination of equations (2.71) and (2.72) readily leads to part (ii) of the claim. In view of equation (2.71), part (i) of the claim will follow upon showing that the random variable $\phi'(\mathbb{G}_{(P,Q)})$ is equal in distribution to $N(0, a\sigma_P^2 + (1-a)\sigma_Q^2)$. In the sequel, write for all $u \in [\delta, 1-\delta]$,

$$w_P(u) = \int_{\mathbb{S}^{d-1}} \frac{w(u,\theta)}{p_\theta(F_\theta^{-1}(u))} d\mu(\theta), \quad w_Q(u) = \int_{\mathbb{S}^{d-1}} \frac{w(u,\theta)}{q_\theta(G_\theta^{-1}(u))} d\mu(\theta).$$

Notice that

$$\begin{aligned}&\phi'(\mathbb{G}_{(P,Q)})\\ &= \int_{\mathbb{S}^{d-1}} \int_\delta^{1-\delta} w(u,\theta)\left(\frac{\sqrt{1-a}\,\mathbb{G}_Q(G_\theta^{-1}(u),\theta)}{q_\theta(G_\theta^{-1}(u))} - \frac{\sqrt{a}\,\mathbb{G}_P(F_\theta^{-1}(u),\theta)}{p_\theta(F_\theta^{-1}(u))}\right) du d\mu(\theta)\\ &= \int_{\mathbb{S}^{d-1}} \int_\delta^{1-\delta} w(u,\theta)\left(\frac{\sqrt{1-a}\,\overline{\mathbb{G}}(u)}{q_\theta(G_\theta^{-1}(u))} - \frac{\sqrt{a}\,\mathbb{G}(u)}{p_\theta(F_\theta^{-1}(u))}\right) du d\mu(\theta)\\ &= \sqrt{1-a}\int_\delta^{1-\delta} w_Q(u)\overline{\mathbb{G}}(u) du - \sqrt{a}\int_\delta^{1-\delta} w_P(u)\mathbb{G}(u)du,\end{aligned}$$

where, on the final line, the interchange of order of integration is valid $\mathbb{P}$-almost surely. Indeed, the sample paths of $\mathbb{G}$ and $\overline{\mathbb{G}}$ are almost surely continuous, whence bounded over $[\delta, 1-\delta]$, which in turn implies that the functions $w(u,\theta)\mathbb{G}(u)/p_\theta(F_\theta^{-1}(u))$ and $w(u,\theta)\overline{\mathbb{G}}(u)/q_\theta(G_\theta^{-1}(u))$ are almost surely bounded, due to the assumptions placed on P and Q. We now make use of the following fact, which is proven below for completeness.

Lemma 16. *Let $f:[\delta,1-\delta]\to\mathbb{R}$ be a Lebesgue-measurable and bounded function. Then, the random variable $\int_\delta^{1-\delta} f(u)\mathbb{G}(u)du$ has Gaussian distribution with mean zero and variance*

$$\mathrm{Var}\left[\int_\delta^{1-\delta} f(u)\mathbb{G}(u)du\right] = \int_0^{1-\delta}\left(\int_{\delta\vee t}^{1-\delta} f(u)du\right)^2 dt - \left(\int_0^{1-\delta}\int_{\delta\vee t}^{1-\delta} f(u)dudt\right)^2.$$

By Lemma 16 and the independence of $\mathbb{G}$ and $\overline{\mathbb{G}}$, we obtain that $\phi'(\mathbb{G}_{(P,Q)})$ has Gaussian distribution with mean zero and variance

$$\begin{aligned}
&\mathrm{Var}\left[\phi'(\mathbb{G}_{(P,Q)})\right] \\
&= a\left[\int_0^{1-\delta}\left(\int_{\delta\vee t}^{1-\delta} w_P(u)du\right)^2 dt - \left(\int_0^{1-\delta}\int_{\delta\vee t}^{1-\delta} w_P(u)dudt\right)^2\right] \\
&+ (1-a)\left[\int_0^{1-\delta}\left(\int_{\delta\vee t}^{1-\delta} w_Q(u)du\right)^2 dt - \left(\int_0^{1-\delta}\int_{\delta\vee t}^{1-\delta} w_Q(u)dudt\right)^2\right].
\end{aligned}$$

Finally, notice that

$$\begin{aligned}
\int_{\delta\vee t}^{1-\delta} w_P(u)du &= \int_{\delta\vee t}^{1-\delta}\int_{\mathbb{S}^{d-1}} \frac{w(u,\theta)}{p_\theta(F_\theta^{-1}(u))}d\mu(\theta)du \\
&= \int_{\mathbb{S}^{d-1}}\int_{\delta\vee t}^{1-\delta} \frac{w(u,\theta)}{p_\theta(F_\theta^{-1}(u))}dud\mu(\theta) \\
&= \int_{\mathbb{S}^{d-1}}\int_{F_\theta^{-1}(\delta\vee t)}^{F_\theta^{-1}(1-\delta)} w(F_\theta(x),\theta)dxd\mu(\theta),
\end{aligned}$$

where, once again, the interchange of the order of integration is valid due to the uniform boundedness of the integrands almost everywhere. A similar computation holds with w_P replaced by w_Q, thus $\mathrm{Var}[\phi'(\mathbb{G}_{(P,Q)})] = a\sigma_P^2 + (1-a)\sigma_Q^2$, and the claim follows. □

It remains to prove Lemmas 14–16.

2.F.1 Proof of Lemma 14

Let $u\in[\delta,1-\delta]$. We prove the claim for F, and an identical argument may then be used for G. We use the abbreviations in equation (2.68). Let $\epsilon>0$ be an arbitrary real number satisfying

$$\epsilon < \inf_{\theta\in\mathbb{S}^{d-1}}\left[F^{-1}(1-\delta/2,\theta) - F^{-1}(1-\delta,\theta)\right].$$

The infimum on the right-hand side is strictly positive by Lemma 15(ii), whose proof below is a consequence of the uniform integrability of $\{p_\theta\}_{\theta\in\mathbb{S}^{d-1}}$, and does not require the present result. By definition of ϵ, we have $\xi_{\theta,u}, \xi_{\theta,u}+\epsilon \in I_\theta$, thus, by absolute continuity of $F(\cdot,\theta)$,

$$\begin{aligned}\inf_{\theta\in\mathbb{S}^{d-1}} \Big[F(\xi_{\theta,u}+\epsilon,\theta) - u\Big] &= \inf_{\theta\in\mathbb{S}^{d-1}} \int_{\xi_{\theta,u}}^{\xi_{\theta,u}+\epsilon} p_\theta(x)dx \\ &\geq \epsilon \inf_{\theta\in\mathbb{S}^{d-1}} \operatorname*{essinf}_{\xi_{\theta,u}\leq x\leq \xi_{\theta,u}+\epsilon} p_\theta(x) \\ &\geq \epsilon \inf_{\theta\in\mathbb{S}^{d-1}} \operatorname*{essinf}_{x\in I_\theta} p_\theta(x) > 0,\end{aligned}$$

where the strict inequality follows from the fact that $\sup_\theta J_{\infty,\delta/2}(P_\theta) < \infty$. After repeating a symmetric argument, we deduce

$$\sup_{\theta\in\mathbb{S}^{d-1}} F(\xi_{\theta,u}-\epsilon,\theta) < u < \inf_{\theta\in\mathbb{S}^{d-1}} F(\xi_{\theta,u}+\epsilon,\theta). \tag{2.73}$$

On the other hand, by definition of quantile, we have for all $\epsilon > 0$ that,

$$\sup_{\theta\in\mathbb{S}^{d-1}} (F+t_kH_{1k})(\xi_{\theta,uk}-\epsilon,\theta) \leq u \leq \inf_{\theta\in\mathbb{S}^{d-1}} (F+t_kH_{1k})(\xi_{\theta,uk},\theta).$$

Recall that H_{1k} converges in $\ell^\infty(\mathcal{H})$ to $H_1 \in \ell^\infty(\mathcal{H})$, thus there exists $C > 0$ such that

$$\sup_{\theta\in\mathbb{S}^{d-1}} F(\xi_{\theta,uk}-\epsilon,\theta) - Ct_k \leq u \leq \inf_{\theta\in\mathbb{S}^{d-1}} F(\xi_{\theta,uk},\theta) + Ct_k.$$

Since $t_k \downarrow 0$, the above display contradicts equation (2.73) for all large enough k if $\xi_{\theta,uk} < \xi_{\theta,u}-\epsilon$, or if $\xi_{\theta,uk} > \xi_{\theta,u}+\epsilon$. We have therefore shown that for all $\epsilon > 0$ small enough, there exists a sufficiently large $K \geq 1$ such that for all $k \geq K$, $\sup_\theta |\xi_{\theta,uk}-\xi_{\theta,u}| \leq \epsilon$, thus leading to the claim. □

2.F.2 Proof of Lemma 15

Consider the family $\mathcal{F} = \{p_\theta\}_{\theta\in\mathbb{S}^{d-1}}$. $\mathcal{F}$ is assumed to be uniformly integrable with respect to the Lebesgue measure over $\mathbb{R}$, and hence is also uniformly integrable with respect to the finite measure $\nu = \lambda|_I$ (that is, the restriction of the Lebesgue measure to the bounded interval I). Since ν does not possess any atoms, uniform integrability of $\mathcal{F}$ is equivalent to $\mathcal{F}$ having uniformly absolutely continuous integrals, by Proposition 4.5.3 of Bogachev (2007). Thus, for any $t > 0$, there exists $\epsilon(t) > 0$ such that for any interval $[\alpha,\beta] \subseteq I$ for which $\beta-\alpha \leq \epsilon(t)$, we have

$$\sup_{\theta\in\mathbb{S}^{d-1}} \big|F(\alpha,\theta) - F(\beta,\theta)\big| \leq t.$$

This proves property (i). For part (ii), choose $t < \delta/4$ and fix the corresponding value of $\epsilon(t)$. Let $u \in [\delta/2, 1-\delta/4]$, and choose a sequence $\theta_j \in \mathbb{S}^{d-1}$ such that for all $j \geq 1$,

$$\big|F^{-1}(u+\delta/4,\theta_j) - F^{-1}(u,\theta_j)\big| \leq \frac{1}{j} + \inf_{\theta\in\mathbb{S}^{d-1}} \big|F^{-1}(u+\delta/4,\theta) - F^{-1}(u,\theta)\big|. \tag{2.74}$$

Let $\alpha_j = F^{-1}(u,\theta_j)$ and $\beta_j = F^{-1}(u+\delta/4,\theta_j)$. Clearly, $F(\beta_j,\theta_j) - F(\alpha_j,\theta_j) = \delta/4 > t$, thus from the uniform absolute continuity of $\{F_\theta\}$ in part (i), it must hold that

$$\beta_j - \alpha_j = F^{-1}(u+\delta/4,\theta_j) - F^{-1}(u,\theta_j) > \epsilon(t).$$

Since this property holds for all $j \geq 1$, we deduce from equation (2.74) that

$$\inf_{\theta \in \mathbb{S}^{d-1}} \left| F^{-1}(u + \delta/4, \theta) - F^{-1}(u, \theta) \right| \geq \epsilon(t)/2.$$

Since $\epsilon(t)$ did not depend on u, we finally arrive at

$$\gamma = \inf_{\delta/2 \leq u \leq 1-\delta/4} \inf_{\theta \in \mathbb{S}^{d-1}} \left| F^{-1}(u + \delta/4, \theta) - F^{-1}(u, \theta) \right| \geq \epsilon(t)/2 > 0,$$

which proves the second claim. To prove the third and fourth, notice simply that

$$\begin{aligned}
&(F + t_k H_{1k})^{-1}(\delta, \theta) \\
&\quad \geq F^{-1}(\delta, \theta) - \sup_{\theta \in \mathbb{S}^{d-1}} \left| F^{-1}(\delta, \theta) - (F + t_k H_{1k})^{-1}(\delta, \theta) \right| \\
&\quad \geq F^{-1}(3\delta/4, \theta) + \gamma - \sup_{\theta \in \mathbb{S}^{d-1}} \left| F^{-1}(\delta, \theta) - (F + t_k H_{1k})^{-1}(\delta, \theta) \right|.
\end{aligned}$$

By Lemma 14, recall that there exists $K \geq 1$ such that $\sup_{\theta \in \mathbb{S}^{d-1}} \left| F^{-1}(\delta, \theta) - (F + t_k H_{1k})^{-1}(\delta, \theta) \right| \leq \gamma$, thus for all such k, we have

$$(F + t_k H_{1k})^{-1}(\delta, \theta) \geq F^{-1}(3\delta/4, \theta).$$

Similarly, up to modifying the value of K, we have for all $k \geq K$,

$$(F + t_k H_{1k})^{-1}(1 - \delta, \theta) \leq F^{-1}(1 - \delta/4, \theta). \tag{2.75}$$

The claim follows from here. □

2.F.3 Proof of Lemma 16

We first prove the claim for step functions f. Let $M \geq 1$ be an integer and define a partition of $[\delta, 1-\delta]$ via $\delta = s_0 < \cdots < s_{M+1} = 1-\delta$. Let $\alpha_0, \dots, \alpha_M \in \mathbb{R}$ and set $f = \sum_{i=0}^{M} \alpha_i I_{[s_i, s_{i+1})}$. Clearly, we may always rewrite f in terms of any refinement of the partition $s_0, \dots, s_{M+1}$. Indeed, for any $K \geq 1$ and any set of real numbers $0 = t_0 < \cdots < t_{K+1} = 1 - \delta$ containing $\{s_0, \dots, s_{M+1}\}$, we may find real numbers $a_0, \dots, a_K$ contained in $\{\alpha_0, \dots, \alpha_M\}$ such that $f = \sum_{k=0}^{K} a_k I_{[t_k, t_{k+1})}$. We must have $a_0 = 0$ when $t_0 = 0 < \delta$. Since $\mathbb{G}$ is almost surely continuous over $[\delta, 1 - \delta]$, and the function f is piecewise continuous, $f\mathbb{G}$ is almost surely Riemann integrable. Therefore, for any choice of the partition $\{t_0, \dots, t_{K+1}\}$ with vanishing mesh as $K \to \infty$, we have

$$\int_{\delta}^{1-\delta} f(u)\mathbb{G}(u)du = \lim_{K \to \infty} \sum_{k=1}^{K+1} \mathbb{G}(t_{k-1})(t_k - t_{k-1})a_{k-1}. \tag{2.76}$$

Notice that,

$$\sum_{k=1}^{K+1} \mathbb{G}(t_{k-1})(t_k - t_{k-1})a_{k-1}$$

$$
\begin{aligned}
&= \sum_{k=1}^{K+1} \mathbb{G}(t_{k-1}) \left[\sum_{j=k-1}^{K} a_j(t_{j+1}-t_j) - \sum_{j=k}^{K} a_j(t_{j+1}-t_j) \right] \\
&= \sum_{k=0}^{K} \mathbb{G}(t_k) \sum_{j=k}^{K} a_j(t_{j+1}-t_j) - \sum_{k=1}^{K+1} \mathbb{G}(t_{k-1}) \sum_{j=k}^{K} a_j(t_{j+1}-t_j) \\
&= \sum_{k=1}^{K} \mathbb{G}(t_k) \sum_{j=k}^{K} a_j(t_{j+1}-t_j) - \sum_{k=1}^{K} \mathbb{G}(t_{k-1}) \sum_{j=k}^{K} a_j(t_{j+1}-t_j) \\
&= \sum_{k=1}^{K} (\mathbb{G}(t_k) - \mathbb{G}(t_{k-1})) \sum_{j=k}^{K} a_j(t_{j+1}-t_j) \\
&= \sum_{k=1}^{K} (\mathbb{G}(t_k) - \mathbb{G}(t_{k-1})) \int_{t_k}^{1-\delta} f(x)dx.
\end{aligned}
$$

Now, since f is bounded, the function $t \in [0, 1-\delta] \mapsto \int_{\delta \vee t}^{1-\delta} f(x)dx$ is continuous and bounded, thus for any partition as in equation (2.76), we obtain

$$
\begin{aligned}
&\int_{\delta}^{1-\delta} f(u)\mathbb{G}(u)du \\
&\quad = \lim_{K\to\infty} \sum_{k=1}^{K+1} (\mathbb{G}(t_k) - \mathbb{G}(t_{k-1})) \int_{t_k}^{1-\delta} f(x)dx = \int_0^{1-\delta} \int_{\delta\vee t}^{1-\delta} f(x)dxd\mathbb{G}(t),
\end{aligned}
$$

where convergence in the final equality is understood to hold in probability (see, for instance, Proposition 2.13 of Revuz and Yor (2013)). It now follows from Proposition 2.2.1 of Denker (1985) that the random variable on the right-hand side of the above display has mean-zero Gaussian distribution, with variance

$$
\int_0^{1-\delta} \left(\int_{\delta\vee t}^{1-\delta} f(u)du \right)^2 dt - \left(\int_0^{1-\delta} \int_{\delta\vee t}^{1-\delta} f(u)dudt \right)^2,
$$

which leads to the claim when f is a step function.

Assume now that f is a Lebesgue measurable bounded function. By Theorem 4.3 of Stein and Shakarchi (2009), there exists a sequence of step functions f_n converging pointwise to f, Lebesgue-almost everywhere on $[\delta, 1-\delta]$. In view of the preceding result, we have in probability,

$$
\begin{aligned}
&\left| \int_{\delta}^{1-\delta} f(u)\mathbb{G}(u)du - \int_0^{1-\delta} \int_{t\vee\delta}^{1-\delta} f(u)dud\mathbb{G}(t) \right| \\
&\qquad\qquad \leq \int_{\delta}^{1-\delta} |f_n - f||\mathbb{G}| + \int_0^{1-\delta} \left(\int_{\delta\vee t}^{1-\delta} |f_n - f| \right) d\mathbb{G}(t).
\end{aligned}
$$

Since f is bounded, we may clearly take the functions f_n to be uniformly bounded. Furthermore, since $\mathbb{G}$ is $\mathbb{P}$-almost surely bounded on the comapct set $[\delta, 1-\delta]$, the first term on the right-hand side of the above display vanishes by the Dominated Convergence Theorem, while the second vanishes in probability by Theorem 2.12 of Revuz and Yor (2013). Deduce that the identity $\int f(u)\mathbb{G}(u)du = \int_0^{1-\delta}\int_{t\vee\delta}^{1-\delta} f(u)dud\mathbb{G}(t)$ holds, with convergence holding in probability. The claim then follows as before, by Denker (1985). □

2.G Proofs of Additional Results

2.G.1 Proof of Proposition 7

Given $\theta \sim \mu$, we have

$$\begin{aligned} W_r^r(P_\theta, Q_\theta) &= \frac{1}{1-2\delta}\int_\delta^{1-\delta} \left|F_\theta^{-1}(u) - G_\theta^{-1}(u)\right|^r du \\ &\lesssim \max_{a\in\{\delta,1-\delta\}} \left|F_\theta^{-1}(a)\right|^r + \max_{a\in\{\delta,1-\delta\}} \left|G_\theta^{-1}(a)\right|^r. \end{aligned}$$

Therefore, it follows from Lemma 3 that

$$\sup_{P,Q\in\mathcal{K}_{2r}(b)} \mathrm{Var}_\mu\left[W_{r,\delta}^r(P_\theta, Q_\theta)\right] \le \sup_{P,Q\in\mathcal{K}_{2r}(b)} \int_{\mathbb{S}^{d-1}} W_r^{2r}(P_\theta, Q_\theta)d\mu(\theta) \le b(4/\delta)^r.$$

Thus, denoting $S_N = \mathrm{SW}_{r,\delta}^{(N)}(P,Q)$, $S = \mathrm{SW}_{r,\delta}(P,Q)$ and $\Delta_N = M_N/\sqrt{N}$, we obtain

$$\begin{aligned} &\mathbb{P}\Big(S \notin \overline{C}_{nm}^{(N)}\Big) \\ &= \mathbb{P}\Big(S \notin \overline{C}_{nm}^{(N)}, |S^r - S_N^r| > \Delta_N\Big) + \mathbb{P}\Big(S \notin \overline{C}_{nm}^{(N)}, |S^r - S_N^r| \le \Delta_N\Big) \\ &\le \mathbb{P}\Big(|S^r - S_N^r| > \Delta_N\Big) \\ &+ \mathbb{P}\Big(\big(\{S^r \le L_{N,nm} - \Delta_N\} \cup \{U_{N,nm} + \Delta_N \le S^r\}\big) \cap \{|S^r - S_N^r| \le \Delta_N\}\Big) \\ &\le \mathbb{P}\Big(|S^r - S_N^r| > \Delta_N\Big) + \mathbb{P}\Big(S_N^r \notin C_{nm}^{(N)}\Big) \\ &= \mathbb{P}\Big(|S^r - S_N^r| > \Delta_N\Big) + \mathbb{E}\left[\mathbb{P}\Big(S_N \notin C_{nm}^{(N)} \,\Big|\, \theta_1, \dots, \theta_N\Big)\right] \\ &\le \frac{\mathrm{Var}_{\theta\sim\mu}\left[W_{r,\delta}^r(P_\theta, Q_\theta)\right]}{N\Delta_N^2} + \alpha \le \frac{bc/\delta^r}{M_N^2} + \alpha, \end{aligned}$$

for a constant $c > 0$ depending only on r, as claimed. □

2.G.2 Proof of Corollary 1

Under the stated conditions on ϵ, δ, it can be seen by direct verification that conditions **B1**–**B3** and the remaining conditions of Theorem 7 hold for the confidence bands of Examples 1–2, for constants K_1, K_2 possibly depending on α. In what follows, the symbol "$\lesssim$" is used to hide constants possibly depending on α, b, δ_0 and r.

Suppose first that $\mathrm{SJ}_{r,\frac{\delta}{2}}(P) \vee \mathrm{SJ}_{r,\frac{\delta}{2}}(Q) = \infty$. Then, we use the bound

$$\begin{aligned}\psi_{\varepsilon,nm} &\lesssim \big(\mathrm{SW}_{\infty,\delta}^{(r-1)}(P,Q) + U_{\varepsilon,n}(P) + U_{\varepsilon,m}(Q)\big)\frac{\kappa_{\varepsilon,n}}{\sqrt{\delta}} \\ &\lesssim \left(\max_{a\in\{\delta/2,1-\delta/2\}} \sup_{\mathbb{S}^{d-1}} \left|F_\theta^{-1}(a)\right|^{r-1}\right)\frac{\kappa_{\varepsilon,n}}{\sqrt{\delta}} \lesssim \frac{\kappa_{\varepsilon,n}}{\delta^{r/2}},\end{aligned}$$

by Lemma 3, under the assumption $P, Q \in \overline{\mathcal{K}}_2(b)$. A similar argument can be used to bound $\varphi_{\varepsilon,nm}$, leading to

$$\begin{aligned}\lambda(C_{nm}^{(N)}) &\leq \left\{\mathrm{SW}_{r,\delta}^r(P,Q) + c\big(\psi_{\varepsilon,nm} + \varphi_{\varepsilon,nm} + \varkappa_N\big)\right\}^{1/r} - \mathrm{SW}_{r,\delta}(P,Q) \\ &\leq c^{1/r}\big(\psi_{\varepsilon,nm} + \varphi_{\varepsilon,nm} + \varkappa_N\big)^{1/r} \\ &\lesssim \varkappa_N^{1/r} + \frac{1}{\sqrt{\delta}}\left(\kappa_{\varepsilon,n}^{1/r} + \kappa_{\varepsilon,m}^{1/r}\right),\end{aligned}$$

with probability at least $1-\epsilon$. Parts (i) and (ii) now follow from Lemma 1 in the case $\mathrm{SJ}_{r,\frac{\delta}{2}}(P) \vee \mathrm{SJ}_{r,\frac{\delta}{2}}(Q) = \infty$.

Suppose now that $\mathrm{SJ}_{r,\frac{\delta}{2}}(P) \vee \mathrm{SJ}_{r,\frac{\delta}{2}}(Q) < \infty$. Using the shorthand notations

$$\Delta = V_{\varepsilon,n}(P) + V_{\varepsilon,m}(Q), \quad S = \mathrm{SW}_{r,\delta}(P,Q),$$

we have the bounds $\psi_{\varepsilon,nm}, \varphi_{\varepsilon,nm} \lesssim (S^r + \Delta)^{\frac{r-1}{r}}\Delta^{\frac{1}{r}}$, whence

$$\lambda(C_{nm}^{(N)}) \lesssim \varkappa_N^{\frac{1}{r}} + \left\{S^r + (S^r+\Delta)^{\frac{r-1}{r}}\Delta^{\frac{1}{r}}\right\}^{\frac{1}{r}} - S,$$

with probability at least $1-\epsilon$. If $S^r \lesssim \Delta$, the right-hand side of the above display is clearly of order $\varkappa_N^{1/r} + \Delta^{1/r}$. Likewise, if $S^r \gtrsim \Delta$, we obtain similarly as in equation (2.41),

$$\begin{aligned}\lambda(C_{nm}^{(N)}) &\lesssim \varkappa_N^{\frac{1}{r}} + S\left[\left\{1 + (1 + (\Delta/S^r))^{\frac{r-1}{r}}(\Delta/S^r)^{\frac{1}{r}}\right\}^{\frac{1}{r}} - 1\right] \\ &\leq \varkappa_N^{\frac{1}{r}} + S(1 + (\Delta/S^r))^{\frac{r-1}{r}}(\Delta/S^r)^{\frac{1}{r}} \\ &\lesssim \varkappa_N^{\frac{1}{r}} + \Delta^{\frac{1}{r}},\end{aligned}$$

with probability at least $1-\epsilon$. The conclusion of the above display thus holds irrespective of S and Δ. The claim now follows by substituting the expressions for $V_{\varepsilon,n}(P)$ stated in Lemma 1 for each of parts (i) and (ii). □

2.G.3 Proof of Corollary 2

We reason similarly as in the proof of Corollary 1. By Theorem 7, we have with probability at least $1-\epsilon$,

$$\lambda(C_{nm}) \leq \left\{\mathrm{SW}_{r,\delta}^r(P,Q) + c\big(\psi_{\varepsilon,nm} + \varphi_{\varepsilon,nm} + \varkappa_N\big)\right\}^{1/r} - \mathrm{SW}_{r,\delta}(P,Q)$$

$$\begin{aligned}&\lesssim \mathrm{SW}_{r,\delta}^{1-r}(P,Q)\big(\psi_{\varepsilon,nm}+\varphi_{\varepsilon,nm}+\varkappa_N\big)\\ &\le \Gamma^{1-r}\big(\psi_{\varepsilon,nm}+\varphi_{\varepsilon,nm}+\varkappa_N\big),\end{aligned}$$

since $\mathrm{SW}_{r,\delta}(P,Q)\geq\Gamma$. The claim now follows by invoking similar bounds on $\psi_{\varepsilon,nm}$ and $\varphi_{\varepsilon,nm}$ as in the proof of Corollary 1. □

2.G.4 Proof of Proposition 8

Notice that $\sigma_P,\sigma_Q>0$ whenever $\mathrm{SW}_{r,\delta}(P,Q)>0$. In view of Theorem 8, it is then a standard result that the percentile bootstrap interval C_{nm}^* satisfies

$$\liminf_{n,m\to\infty}\mathbb{P}(\mathrm{SW}_{r,\delta}(P,Q)\in C_{nm}^*)\geq 1-\alpha/2, \tag{2.77}$$

under the assumptions of Theorem 8 and the assumption $\mathrm{SW}_{r,\delta}(P,Q)>0$ (see, for instance, Lemma 23.3 of van der Vaart (1998)). Therefore, when these assumptions hold, we have

$$\begin{aligned}&\mathbb{P}(\mathrm{SW}_{r,\delta}(P,Q)\notin C_{nm})\\ &\quad=\mathbb{P}(\mathrm{SW}_{r,\delta}(P,Q)\notin C_{nm}^*,0\notin C_{nm}^\dagger)+\mathbb{P}(\mathrm{SW}_{r,\delta}(P,Q)\notin C_{nm}^\dagger,0\in C_{nm}^\dagger)\\ &\quad\leq\mathbb{P}(\mathrm{SW}_{r,\delta}(P,Q)\notin C_{nm}^*)+\mathbb{P}(\mathrm{SW}_{r,\delta}(P,Q)\notin C_{nm}^\dagger)\leq\alpha+o(1),\end{aligned}$$

where on the final line, we use equation (2.77) and Proposition 7. Notice that the assumptions of Proposition 7 are satisfied, since $\overline{\mathcal{K}}_2\subseteq\mathcal{K}_{2r}$. On the other hand, when $\mathrm{SW}_{r,\delta}(P,Q)=0$, we obtain

$$\begin{aligned}&\mathbb{P}(\mathrm{SW}_{r,\delta}(P,Q)\notin C_{nm})\\ &\quad=\mathbb{P}(\mathrm{SW}_{r,\delta}(P,Q)\notin C_{nm}^*,0\notin C_{nm}^\dagger)+\mathbb{P}(\mathrm{SW}_{r,\delta}(P,Q)\notin C_{nm}^\dagger,0\in C_{nm}^\dagger)\\ &\quad\leq 2\mathbb{P}(0\notin C_{nm}^\dagger)\leq\alpha+o(1).\end{aligned}$$

This proves the stated asymptotic coverage property of the confidence interval C_{nm}. In order to bound its length, note that it is a direct consequence of Theorem 8 that the bootstrap quantiles $F_{nm}^*(\alpha/2)$ and $F_{nm}^*(1-\alpha/2)$ are of the order $O_p(n^{-1/2})$ as $n/(n+m)\to a\in(0,1)$. Thus,

$$b_{nm}^*-a_{nm}^*=O_p(n^{-1/2}),$$

where we write $C_{nm}^*=[(a_{nm}^*)^{1/r},(b_{nm}^*)^{1/r}]$. Furthermore, under the conditions of Theorem 8, and for $N\asymp n^{r^2}$ and $M_N\asymp\log N$, it can be deduced from Theorem 7 and Corollary 1 that when $\mathrm{SW}_{r,\delta}(P,Q)=0$, we have

$$b_{nm}^\dagger-a_{nm}^\dagger=O_p\left(\left(\frac{\log n}{n}\right)^{\frac{r}{2}}\right)$$

where we write $C_{nm}^\dagger=[(a_{nm}^\dagger)^{1/r},(b_{nm}^\dagger)^{1/r}]$. Finally, as in the above proof of coverage of C_{nm}, when $\mathrm{SW}_{r,\delta}(P,Q)>0$ we have $C_{nm}=C_{nm}^*$ with probability at least $1-\alpha/2-o(1)$, while when $\mathrm{SW}_{r,\delta}(P,Q)=0$ we have $C_{nm}=C_{nm}^\dagger$ with probability at least $1-\alpha/2-o(1)$.

Combine these facts to deduce that for any $\epsilon > 0$, there exist constants $C, n_0 > 0$ such that for all $n \geq n_0$,

$$b_{nm} - a_{nm} \leq C \left(\left(\frac{\log n}{n} \right)^{\frac{r}{2}} + \frac{\mathrm{SW}_{r,\delta}(P,Q)}{\sqrt{n}} \right)$$

with probability at least $1 - \alpha/2 - \epsilon$. Choosing $\epsilon = \alpha/2$ leads to the claim. □

The proof is straightforward. We have,

$$\begin{aligned} \mathbb{P}\Big(\eta_0 \notin \overline{C}_{nm}^{(N)}\Big) &\leq \mathbb{P}\Big(\bar{\ell}_{nm}^{(N)}(\eta_0) > \epsilon\Big) \\ &\leq \mathbb{P}\Big(\bar{\ell}_{nm}^{(N)}(\eta_0) > \epsilon_0\Big) = \mathbb{P}\Big(\bar{\ell}_{nm}^{(N)}(\eta_0) > \mathrm{SW}_{r,\delta}(P, P_{\eta_0})\Big). \end{aligned}$$

The claim now follows from Proposition 7. □

2.G.5 Proof of Example 2

We begin by proving the validity of the inequality in equation (2.23). Let $\mathcal{A}$ be a collection of sets, and let $\mathcal{S}_{\mathcal{A}}(n)$ denote the shattering number (Vapnik, 2013) of $\mathcal{A}$. The relative VC inequality is then given by

$$\mathbb{P}\left(\sup_{A \in \mathcal{A}} \frac{|P_n(A) - P(A)|}{\sqrt{P_n(A)}} \geq t \right) \leq 4\mathcal{S}_{\mathcal{A}}(2n) e^{-nt^2/4}, \quad t > 0.$$

Letting $\mathcal{A} = \{(-\infty, x] : x \in \mathbb{R}\}$ and $\mathcal{A} = \{[x, \infty) : x \in \mathbb{R}\}$ respectively, we obtain

$$\begin{aligned} \mathbb{P}\left(\sup_{x \in \mathbb{R}} \frac{|F_n(x) - F(x)|}{\sqrt{F_n(x)}} \geq t \right) &\leq 4(2n+1)e^{-nt^2/4}, \\ \mathbb{P}\left(\sup_{x \in \mathbb{R}} \frac{|F_n(x) - F(x)|}{\sqrt{1 - F_n(x)}} \geq t \right) &\leq 4(2n+1)e^{-nt^2/4}, \end{aligned}$$

for all $t > 0$. By a union bound and the fact that $u(1-u) \geq \frac{1}{2}(u \wedge (1-u))$ for all $u \in [0,1]$, we arrive at

$$\mathbb{P}\left(\sup_{x \in \mathbb{R}} \frac{|F_n(x) - F(x)|}{\sqrt{F_n(x)(1 - F_n(x))}} \geq t \right) \leq 8(2n+1)e^{-\frac{nt^2}{16}}.$$

Setting $t = \nu_{\alpha,n} := \sqrt{\frac{16}{n}\left[\log(16/\alpha) + \log(2n+1)\right]}$, we deduce that with probability at least $1 - \alpha/2$,

$$|F_n(x) - F(x)| \leq \nu_{\alpha,n} \sqrt{F_n(x)(1 - F_n(x))}, \quad \forall x \in \mathbb{R}. \tag{2.78}$$

This proves the validity of equation (2.23).

We now invert equation (2.78) to obtain the functions $\gamma_{\alpha,n}$ and $\eta_{\alpha,n}$ which lead to a quantile confidence band. We will require the following definitions of lower CDF and upper quantile function,

$$\overline{F}(x) := \lim_{y \to x^-} F(y) = \mathbb{P}(X_1 < x), \quad \overline{F}^{-1}(u) = \inf\big\{x \in \mathbb{R} : \overline{F}(x) > u\big\},$$

with empirical analogues given by

$$\overline{F}_n(x) := \lim_{y \to x^-} F_n(y) = \frac{1}{n}\sum_{i=1}^n I(X_i < x), \quad \overline{F}_n^{-1}(u) = \inf\big\{x \in \mathbb{R} : \overline{F}_n(x) > u\big\}.$$

Notice that F and $\overline{F}^{-1}$ are right continuous, whereas $\overline{F}$ and F^{-1} are left continuous. Furthermore, we make use of the following elementary inequalities relating quantile functions and CDFs,

$$F_n(x) \geq u \Longrightarrow x \geq F_n^{-1}(u), \tag{2.79}$$

$$\overline{F}_n(x) \leq u \Longrightarrow x \leq \overline{F}_n^{-1}(u), \tag{2.80}$$

$$F(x) \geq u \Longleftrightarrow x \geq F^{-1}(u), \tag{2.81}$$

$$\overline{F}(x) \leq u \Longleftrightarrow x \leq \overline{F}^{-1}(u). \tag{2.82}$$

We now turn to the proof. The calculations which follow are elementary, but tedious. Let $v = F(x)$. By equation (2.78), we have with probability at least $1 - \alpha/2$ that for all $x \in \mathbb{R}$,

$$\begin{aligned}
F_n(x) + \nu_{\alpha,n}\sqrt{F_n(x)(1-F_n(x))} \geq v &\geq F_n(x) - \nu_{\alpha,n}\sqrt{F_n(x)(1-F_n(x))} \\
&\Longrightarrow F_n(x)(1-F_n(x))\nu_{\alpha,n}^2 \geq (v - F_n(x))^2 \\
&\Longrightarrow (F_n(x) - F_n(x)^2)\nu_{\alpha,n}^2 \geq v^2 - 2vF_n(x) + F_n(x)^2 \\
&\Longrightarrow F_n(x)^2(1+\nu_{\alpha,n}^2) - F_n(x)\big(2v + \nu_{\alpha,n}^2\big) + v^2 \leq 0 \\
&\Longrightarrow F_n(x) \geq \frac{2v + \nu_{\alpha,n}^2}{2(1+\nu_{\alpha,n}^2)} - \frac{\sqrt{\big[2v + \nu_{\alpha,n}^2\big]^2 - 4(1+\nu_{\alpha,n}^2)v^2}}{2(1+\nu_{\alpha,n}^2)} \\
&\Longrightarrow F_n(x) \geq \frac{2v + \nu_{\alpha,n}^2}{2(1+\nu_{\alpha,n}^2)} - \frac{\nu_{\alpha,n}\sqrt{\nu_{\alpha,n}^2 + 4v(1-v)}}{2(1+\nu_{\alpha,n}^2)} = \gamma_{\alpha,n}(v) \\
&\Longrightarrow x \geq F_n^{-1}(\gamma_{\alpha,n}(v)), \qquad \text{(By (2.79))} \\
&\Longrightarrow x \geq F_n^{-1}(\gamma_{\alpha,n}(F(x))).
\end{aligned}$$

Now, let $u \in (0,1)$. Setting $x = F^{-1}(u)$ and using the fact that $F \circ F^{-1}(u) \geq u$ by equation (2.81), the above display implies

$$F^{-1}(u) \geq F_n^{-1}(\gamma_{\alpha,n}(F \circ F^{-1}(u))) \geq F_n^{-1}(\gamma_{\alpha,n}(u)),$$

uniformly in $u \in (0,1)$, with probability at least $1 - \alpha/2$.

We now turn to an upper confidence bound on $F^{-1}(u)$. Upon taking limits from the left in equation (2.78), we obtain

$$\overline{F}_n(x) - \nu_{\alpha,n}\sqrt{\overline{F}_n(x)(1-\overline{F}_n(x))} \leq \overline{F}(x) \leq \overline{F}_n(x) + \nu_{\alpha,n}\sqrt{\overline{F}_n(x)(1-\overline{F}_n(x))}$$

uniformly in $x \in \mathbb{R}$, on the same event of probability at least $1 - \alpha/2$. Thus, letting $v = \overline{F}(x)$, we have

$$\nu_{\alpha,n}^2 \overline{F}_n(x)(1-\overline{F}_n(x)) \geq (\overline{F}_n(x) - v)^2$$

$$
\begin{aligned}
&\implies \nu_{\alpha,n}^2 \overline{F}_n(x) - \nu_{\alpha,n}^2 \overline{F}_n(x)^2 \geq \overline{F}_n(x)^2 - 2v\overline{F}_n(x) + v^2 \\
&\implies \overline{F}_n(x)^2(1+\nu_{\alpha,n}^2) - (\nu_{\alpha,n}^2 + 2v)\overline{F}_n(x) + v^2 \leq 0 \\
&\implies \overline{F}_n(x) \leq \frac{\nu_{\alpha,n}^2 + 2v}{2(1+\nu_{\alpha,n}^2)} + \frac{\sqrt{\left[2v + \nu_{\alpha,n}^2\right]^2 - 4(1+\nu_{\alpha,n}^2)v^2}}{2(1+\nu_{\alpha,n}^2)} \\
&\implies \overline{F}_n(x) \leq \frac{\nu_{\alpha,n}^2 + 2v + \nu_{\alpha,n}\sqrt{\nu_{\alpha,n}^2 + 4v(1-v)}}{2(1+\nu_{\alpha,n}^2)} = \eta_{\alpha,n}(v) \\
&\implies x \leq \overline{F}_n^{-1}(\eta_{\alpha,n}(v)), \qquad \text{(By (2.80))} \\
&\implies x \leq \overline{F}_n^{-1}(\eta_{\alpha,n}(\overline{F}(x))).
\end{aligned}
$$

Therefore, setting $x = \overline{F}^{-1}(u)$ for $u \in (0,1)$, and using the fact that $\overline{F} \circ \overline{F}^{-1}(u) \leq u$ by equation (2.82), we obtain

$$\overline{F}^{-1}(u) \leq \overline{F}_n^{-1}\big(\eta_{\alpha,n}(\overline{F}(\overline{F}^{-1}(u))\big) \leq \overline{F}_n^{-1}\big(\eta_{\alpha,n}(u)\big).$$

Upon taking limits from the right, this implies

$$F^{-1}(u) \leq F_n^{-1}\big(\eta_{\alpha,n}(u)\big),$$

uniformly in u with probability at least $1-\alpha/2$. We conclude that

$$\mathbb{P}\Big(F_n^{-1}\big(\gamma_{\alpha,n}(u)\big) \leq F^{-1}(u) \leq F_n^{-1}\big(\eta_{\alpha,n}(u)\big),\ \forall u \in (0,1)\Big) \geq 1-\alpha/2.$$

The validity of equation (2.24) follows. □

Chapter 3

Sharp Rates for Empirical Optimal Transport with Smooth Costs

3.1 Introduction

In this chapter, we return our attention to the study of the multidimensional Wasserstein distance, and more generally, to a broad class of optimal transport divergence functionals $\mathcal{T}_c$. Our goal will be to characterize the expected convergence rate of empirical estimators of such optimal transport costs. Though this question has already been studied in great generality in the literature, our aim in this chapter will be to highlight some unexpected phenomena that arise when the cost function is smooth.

For concreteness, we focus throughout on the optimal transport problem over the Euclidean space $\mathbb{R}^d$ for some integer $d \geq 1$. Let $\mathcal{X}, \mathcal{Y} \subseteq \mathbb{R}^d$, and $P \in \mathcal{P}(\mathcal{X})$ and $Q \in \mathcal{P}(\mathcal{Y})$. Given a nonnegative cost function $c : \mathcal{X} \times \mathcal{Y} \to \mathbb{R}$, recall that the *optimal transport cost* based on c is defined by

$$\mathcal{T}_c(P, Q) = \inf_{\pi \in \Pi(P,Q)} \int c(x, y) d\pi(x, y),$$

where $\Pi(P, Q)$ denotes the set of *couplings* between P and Q. In statistical contexts, the measures P and Q are typically unknown, and it is necessary to estimate the optimal transport cost between them on the basis of i.i.d. observations[1] $X_1, \dots, X_n \sim P$ and $Y_1, \dots, Y_n \sim Q$. A canonical choice is the plugin estimator $\mathcal{T}_c(P_n, Q_n)$, obtained by replacing P and Q by their corresponding empirical measures:

$$P_n = \frac{1}{n} \sum_{i=1}^n \delta_{X_i}, \quad Q_n = \frac{1}{n} \sum_{i=1}^n \delta_{Y_i}.$$

We call this quantity the *empirical optimal transport cost*, and we seek sharp upper and lower bounds on the expected gap between the empirical optimal transport cost and its population

[1] The two sample sizes are assumed to be equal in this chapter for ease of exposition.

counterpart:

$$\Delta_n(c) = \mathbb{E}\big|\mathcal{T}_c(P_n, Q_n) - \mathcal{T}_c(P, Q)\big|. \tag{3.1}$$

We highlight the dependence of Δ_n on c because a key finding of our work is that the rate of decay of Δ_n is driven by properties of the cost, and can improve significantly when c is smooth.

To illustrate the phenomena we have in mind, we turn to perhaps the most widely-used cost functions: those of the form $c_p(x, y) = \|x - y\|^p$, $p \geq 1$, which give rise to the p-Wasserstein distances $W_p = \mathcal{T}_{c_p}^{1/p}$. The convergence rate of the empirical p-Wasserstein distance $W_p(P_n, Q_n)$ to its population counterpart $W_p(P, Q)$ is a well-studied problem; for instance, assuming for simplicity of exposition that $\mathcal{X} = \mathcal{Y}$ is a compact set, Fournier and Guillin (2015) prove that there exists a constant $C_d > 0$, depending only on d and $\mathcal{X}$, such that

$$\mathbb{E}W_p(P_n, P) \leq \big[\mathbb{E}W_p^p(P_n, P)\big]^{\frac{1}{p}} \leq C_d n^{-1/d}, \tag{3.2}$$

whenever $d > 2p$. Since W_p is a metric, it follows that

$$\mathbb{E}\big|W_p(P_n, Q_n) - W_p(P, Q)\big| \leq \mathbb{E}W_p(P_n, P) + \mathbb{E}W_p(Q_n, Q) \leq 2C_d n^{-1/d}. \tag{3.3}$$

The $n^{-1/d}$ rate in equation (3.3) is well known to be inherent to statistical optimal transport problems. In particular, it was shown by Niles-Weed and Rigollet (2022) that, up to polylogarithmic factors, no estimator of $W_p(P, Q)$ improves on the rate in equation (3.3) uniformly over all pairs of measures (P, Q). Nevertheless, one of the main contributions of this chapter is to show that this bound is only tight when $P = Q$, and can otherwise be improved up to quadratically. Indeed, our results imply the bound

$$\Delta_n(c_p) = \mathbb{E}\big|W_p^p(P_n, Q_n) - W_p^p(P, Q)\big| \lesssim \begin{cases} n^{-p/d}, & 1 \leq p \leq 2 \\ n^{-2/d}, & 2 \leq p < \infty, \end{cases} \tag{3.4}$$

which, as we shall see, entails

$$\mathbb{E}\big|W_p(P_n, Q_n) - W_p(P, Q)\big| \lesssim \delta_0^{1-p} \begin{cases} n^{-p/d}, & 1 \leq p \leq 2 \\ n^{-2/d}, & 2 \leq p < \infty \end{cases}, \quad \text{if } W_p(P, Q) \geq \delta_0 > 0. \tag{3.5}$$

Whenever $p > 1$, equations (3.4) and (3.5) provide a significant sharpening of the naive estimate in equation (3.3). We show that such improvements arise due to the Hölder smoothness of the cost c_p, and in fact, similar rates of convergence for $\Delta_n(c)$ are enjoyed by a much broader collection of smooth cost functions. Beyond smoothness assumptions on c, we establish our main results under the following broad structural condition, which is presumed throughout the sequel,

(H0) The cost function $c : \mathcal{X} \times \mathcal{Y} \to \mathbb{R}$ is nonnegative, and takes the form $c(x, y) = h(x - y)$ where $h : \mathbb{R}^d \to \mathbb{R}_+$ is convex, even, and lower semi-continuous.

Before summarizing our main results and comparing them to further existing literature, we begin with an idealized example which illustrates the role of these conditions.

3.1.1 Example: Location Families

For simplicity, we limit this example to upper bounding the following one-sample analogue of $\Delta_n(c)$,

$$\mathbb{E}|\mathcal{T}_c(P_n, Q) - \mathcal{T}_c(P, Q)|.$$

We also continue to assume for simplicity that $\mathcal{X} = \mathcal{Y}$ is a convex and compact set. Let c be any cost function such that condition (**H0**) holds, and assume there exists $\alpha \in (1, 2]$ such that $h \in \mathscr{C}^\alpha(\mathcal{X})$.

Let $P, Q \in \mathcal{P}(\mathcal{X})$ be any two measures differing only by a location transformation with respect to a fixed vector $z_0 \in \mathbb{R}^d$, in the sense that $Q = T_{0\#}P := P(T_0^{-1}(\cdot))$, where $T_0(z) = z + z_0$ and $\#$ denotes the pushforward operator. In this example, it is simple to find an optimal coupling between P and Q. Indeed, recall that h is convex and even under (**H0**), thus for all $\pi \in \Pi(P, Q)$, Jensen's inequality implies

$$\int h(x - y) d\pi(x, y) \geq h\left(\int x dP(x) - \int y dQ(y)\right) = h(z_0).$$

Thus, $\mathcal{T}_c(P, Q) \geq h(z_0)$, and the lower bound is achieved by the coupling $\pi = (Id, T_0)_\# P$, implying that T_0 is an optimal transport map from P to Q. On the other hand, for all couplings $\pi_n \in \Pi(P_n, P)$, $\gamma_n = (Id, T_0)_\# \pi_n$ is a (typically suboptimal) coupling between P_n and Q, whence

$$\mathcal{T}_c(P_n, Q) \leq \int c(z, y) d\gamma_n(z, y) = \int c(z, T_0(x)) d\pi_n(z, x).$$

Since we assumed that the Hölder norm $\Lambda := \|h\|_{\mathscr{C}^\alpha(\mathcal{X})}$ is finite, h is close to its first-order Taylor expansion. Specifically, we obtain from the above display,

$$\mathcal{T}_c(P_n, Q) \leq \int \Big[h(x - T_0(x)) + \langle \nabla h(x - T_0(x)), z - x \rangle + \Lambda \|z - x\|^\alpha \Big] d\pi_n(z, x). \quad (3.6)$$

Due to the marginal constraints in the definition of π_n, equation (3.6) is tantamount to

$$\mathcal{T}_c(P_n, Q) \leq \mathcal{T}_c(P, Q) + \int \langle \nabla h(z_0), \cdot \rangle d(P_n - P) + \Lambda \int \|x - z\|^\alpha d\pi_n(x, z). \quad (3.7)$$

The final term of the above display is manifestly the $\|\cdot\|^\alpha$-transport cost between P_n and P, with respect to a possibly suboptimal coupling $\pi_n \in \Pi(P_n, P)$. Since it holds for any choice of π_n, taking the infimum over such couplings leads to

$$\mathcal{T}_c(P_n, Q) \leq \mathcal{T}_c(P, Q) + \int \langle \nabla h(z_0), \cdot \rangle d(P_n - P) + \Lambda W_\alpha^\alpha(P_n, P). \quad (3.8)$$

The second term on the right-hand side of the above display is a mean-zero sample average, and hence typically decays at the rate $n^{-1/2}$ in probability. Equation (3.8) thus provides an upper bound on $\mathcal{T}_c(P_n, Q) - \mathcal{T}_c(P, Q)$ which is primarily driven by the rate of convergence of the empirical measure under the optimal transport cost with respect to $\|\cdot\|^\alpha$, which we refer to as the α-transport cost in the sequel. By equation (3.2), we arrive at the following one-sided estimate whenever $d \geq 5$,

$$\mathbb{E}\Big[\mathcal{T}_c(P_n, Q) - \mathcal{T}_c(P, Q)\Big] \leq \Lambda \mathbb{E}\big[W_\alpha^\alpha(P_n, P)\big] \leq \Lambda C_d n^{-\alpha/d}. \tag{3.9}$$

Although equation (3.9) does not imply an upper bound in expected absolute value, it captures the main features of our problem; as we shall see, a simple extension of the above derivations leads to the bound

$$\mathbb{E}\big|\mathcal{T}_c(P_n, Q) - \mathcal{T}_c(P, Q)\big| \leq C_0 n^{-\alpha/d}, \quad \text{for all } d \geq 5, \tag{3.10}$$

for a large enough constant $C_0 > 0$ depending on d, $\mathcal{X}$ and Λ. Equation (3.10) shows that, for the class of cost functions under consideration, the rate of convergence of the empirical optimal transport cost is largely driven by the smoothness of c when P and Q differ merely in mean. In particular, notice that the cost $h(x) = \|x\|^p$ satisfies $h \in \mathscr{C}^p$ for all $p \geq 1$, which implies the previously announced result (3.4) in the special case of one-sample location families.

This fast rate of convergence in equation (3.10) arose in the present example because the first-order term in the Taylor expansion (3.6) is negligible, leading to a rate driven only by its remainder. While this argument cannot easily be extended to general measures P and Q, its conclusion turns out to be generic, as we now describe.

3.1.2 Our Contributions

The primary contribution of this chapter is to provide sharp upper and lower bounds on $\Delta_n(c)$ for smooth costs satisfying condition (**H0**). In this setting, our main result informally states that whenever $h \in \mathscr{C}^\alpha$ for some $\alpha > 0$,

$$\Delta_n(c) \lesssim \begin{cases} n^{-\alpha/d}, & 0 \leq \alpha \leq 2 \\ n^{-2/d}, & 2 \leq \alpha < \infty \end{cases}, \qquad \text{for all } d \geq 5. \tag{3.11}$$

This upper bound is stated formally in Theorem 9 under the assumption that P and Q admit bounded support. Under additional conditions on c, we extend this result to measures P and Q with unbounded support, satisfying appropriate tail assumptions, in Theorem 11 and Corollary 5. As in equation (3.5), our results have natural implications for the convergence rate of empirical Wasserstein distances, which we discuss in Corollary 4. In view of Section 3.1.1, the convergence rate (3.11) admits a natural interpretation: the first order term in a formal expansion of the empirical optimal transport cost is typically negligible when $d \geq 5$, leading to a rate that improves with the smoothness parameter $\alpha \in (0, 2)$. When $\alpha \geq 2$, the quadratic term in this expansion is not negligible, thus faster rates do not occur without stronger conditions.

At the heart of our proofs is the Kantorovich dual formulation of the optimal transport problem—summarized in Section 3.1.4—which allows us to reduce the problem of bounding $\Delta_n(c)$ to that of bounding the expected suprema of empirical processes indexed by collections of sufficiently regular Kantorovich potentials. While characterizing the regularity of these potentials is routine when P and Q are compactly supported (Gangbo and McCann (1996), Appendix C), the bulk of our efforts lies in the case where they admit unbounded support. In this setting, one of our key technical contributions is to provide quantitative L^∞ estimates on the displacement induced by the optimal coupling between any two measures satisfying appropriate tail conditions (Theorem 10). For instance, the following is a special case of our result for the p-transport cost.

Theorem (Informal). Let $P, Q \in \mathcal{P}(\mathbb{R}^d)$ and $p > 1$. Let Q be a σ^2-sub-Gaussian measure (Boucheron, Lugosi, and Massart, 2013) and P have finite p-th moment, and assume there exist constants $c_1, c_2 > 0$ such that $P(B_{x,1}) \geq c_1 \exp(-c_2 \|x\|^2)$ for all $x \in \mathbb{R}^d$. Then, for any optimal coupling π between P and Q with respect to the cost $c_p(x,y) = \|x-y\|^p$,

$$\|y\| \lesssim \sigma(\|x\| + 1), \qquad \text{for } \pi\text{-a.e. } (x,y). \tag{3.12}$$

In particular, if there exists an optimal transport map T from P to Q with respect to c_p, then

$$\|T(x)\| \lesssim \sigma(\|x\| + 1), \qquad \text{for } P\text{-a.e. } x.$$

Analogues of equation (3.12) have previously been derived by Colombo and Fathi (2021) in the special case where P is a Gaussian measure and $p = 2$, and we further discuss these results below the statement of Theorem 10. As we shall see, equation (3.12) leads to estimates on the local Lipschitz constants of Kantorovich potentials between any two, possibly atomic probability measures, and forms the basis of our main results when P and Q have unbounded support. These results are quantitative analogues of the fact, proved by Gangbo and McCann (1996), that Kantorovich potentials are locally Lipschitz under mild smoothness conditions on c.

In Section 3.4.1 we explicitly construct measures P and Q for which inequality (3.11) is achieved up to universal constants, inspired by the example in Section 3.1.1. While this result proves that our upper bounds cannot generally be improved, it does not preclude the possibility that there exists another estimator $\widehat{\mathcal{T}}_n$, i.e. a measurable function of $X_1, Y_1, \dots, X_n, Y_n$, for which the quantity $\mathbb{E}|\widehat{\mathcal{T}}_n - \mathcal{T}_c(P,Q)|$ scales at a faster rate than that of equation (3.11), uniformly over pairs of measures P, Q. We prove in Section 3.4.2 that, in an information theoretic sense, such an improvement is not possible up to polylogarithmic factors.

Though we prove inequality (3.11) for all $d \geq 5$, notice that it does not generally hold for all $d \geq 1$. Indeed, it is a simple observation that the empirical optimal transport cost cannot generally achieve a faster rate of convergence than $n^{-1/2}$ (Niles-Weed and Rigollet, 2022). The probabilistic behavior of the empirical costs is therefore qualitatively different in low dimension. While our proof techniques for bounded measures can be extended to the case $d \leq 4$, they do not appear to yield tight results for certain values of $\alpha > 0$; see Remark 1 below. Similarly, our techniques for unbounded measures do not generally appear to be tight in the low-dimensional case. Since our goal in this chapter is to obtain sharp convergence rates, we assume in what follows that $d \geq 5$, where we are able to establish exact results.

Outline of the remainder of the chapter In Sections 3.1.3 and 3.1.4, we review prior work and recall some important preliminary results on the duality theory of transport costs.

Section 3.2 contains our main results for compactly supported measures. In Section 3.3, we extend these results to the unbounded case. Lower bounds appear in Section 3.4. The proofs of certain intermediary results from Sections 3.2–3.4 are respectively deferred to Appendices 3.A–3.C.

3.1.3 Related Work

Upper bounds on the expected deviation $\Delta_n(c)$ are available in the literature for several special cases. The closest to our setting is the quadratic cost $c_2(x, y) = \|x - y\|^2$, for which Chizat et al. (2020) prove that $\Delta_n(c_2) \lesssim n^{-2/d}$ when P and Q are compactly supported. Their proof hinges upon the Knott-Smith optimality criterion, which allows them to relate $\Delta_n(c_2)$ to suprema of empirical processes indexed by convex potentials, which are in fact globally Lipschitz since P and Q are assumed compact. Empirical processes indexed by globally Lipschitz convex functions are well-studied (Bronshtein, 1976; Guntuboyina and Sen, 2012), and lead to their result. Our results extend theirs in two directions: we replace c_2 by any smooth cost, and we remove the condition that the measures be compactly supported. When P and Q are compactly supported and $\alpha = 2$, our proof strategy mirrors that of Chizat et al. (2020): though the potentials arising for other costs are not necessarily convex, it is still possible to use existing empirical process theory bounds to obtain sharp rates. On the other hand, when P and Q have unbounded support, the relevant potentials may not even be globally Lipschitz, and our proof requires significant new techniques.

Faster rates of convergence for estimating optimal transport costs are achievable under strong conditions on $\mathcal{X}$ and $\mathcal{Y}$. For instance, when c is a metric raised to a power $p \geq 1$, the bound $\Delta_n(c) \lesssim n^{-1/2}$ is known to hold when $\mathcal{X}$ and $\mathcal{Y}$ are countable (Sommerfeld and Munk, 2018; Tameling, Sommerfeld, and Munk, 2019) or one-dimensional (Munk and Czado, 1998; Freitag and Munk, 2005; Bobkov and Ledoux, 2019; del Barrio, Gordaliza, and Loubes, 2019) (see also Chapter 2). In both of these cases, the corresponding empirical p-Wasserstein distance is known to exhibit distinct convergence rates depending on whether P and Q are vanishingly close or not, similar to our findings in equation (3.5). While these two examples form important special cases, their underlying proof techniques are closely tied to characterizations of the optimal transport problem which are only available for discrete and one-dimensional measures, and do not shed more general light on the behaviour of $\mathcal{T}_c(P_n, Q_n)$.

Though the naive bound in equation (3.3) is loose for $p > 1$ when $W_p(P, Q)$ is bounded away from zero, Liang (2019) and Niles-Weed and Rigollet (2022) show that it cannot generally be improved by more than a polylogarithmic factor when no separation conditions are placed on P and Q. Recall that this upper bound arose from the convergence rate of P_n under the p-Wasserstein distance in equation (3.2). The study of such convergence rates was initiated by Dudley (1969) in the special case $p = 1$, who also used arguments from empirical process theory, due to the dual characterization of W_1 as a supremum over Lipschitz functions (Villani, 2003). For $p > 1$, distinct techniques have been used to study this problem in great generality by Boissard and Le Gouic (2014); Fournier and Guillin (2015); Bobkov and Ledoux (2019); Weed and Bach (2019); Singh and Póczos (2019); Lei (2020), and references therein. Dudley (1969) also derived deterministic lower bounds on the quality of approximating P by any discrete measure supported on n points under W_1—we build upon these results to obtain our lower bounds on $\Delta_n(c)$ in Section 3.4.1.

Another line of work has sought to understand optimal rates of estimation for Wasserstein distances when the *densities*—rather than the cost—are smooth. These works (Liang, 2021; Singh et al., 2018; Niles-Weed and Berthet, 2022) show that the plugin empirical estimator $W_p(P_n, Q_n)$ for $W_p(P,Q)$ is suboptimal if P and Q have smooth densities, but that replacing P_n and Q_n by appropriate nonparametric density estimators suffices to obtain optimal rates of estimation. Under similar conditions on P and Q, it is also possible to construct appropriate smooth estimators of the optimal map between P and Q (Hütter and Rigollet, 2021). Our work takes a quite different perspective: rather than adding additional conditions on P and Q, we show that the rates of convergence of empirical estimators improve under additional smoothness conditions on the cost.

3.1.4 Notation and Further Background on the Optimal Transport Problem

Our proofs make repeated use of the Kantorovich dual formulation of the optimal transport problem, described in Section 1.3. In what follows, we state several additional properties related to the Kantorovich duality which can be deduced from (Villani, 2008, Chapter 5), and will be used repeatedly throughout this chapter.

Define for all $\phi \in L^1(P)$ and $\psi \in L^1(Q)$ the functional

$$J_{P,Q}(\phi, \psi) = \int \phi dP + \int \psi dQ.$$

The regularity condition (**H0**) is sufficient to imply that the Kantorovich duality

$$\mathcal{T}_c(P,Q) = \sup_{(\phi,\psi)\in\Phi_c(P,Q)} J_{P,Q}(\phi,\psi), \tag{3.13}$$

holds, where we recall that $\Phi_c(P,Q)$ is defined in equation (1.10). If we further assume $c(x,y) \leq c_1(x) + c_2(y)$ for some $c_1 \in L^1(P)$ and $c_2 \in L^1(Q)$, then the supremum in equation (3.13) is achieved by a pair of optimal Kantorovich potentials (ϕ_0, ψ_0), which we recall can always be taken to be of the form

$$\psi_0(y) = \phi_0^c(y) = \inf_{x\in\mathcal{X}} \Big\{ c(x,y) - \phi_0(x) \Big\}, \quad y \in \mathcal{Y}.$$

Recall that when the cost function is taken to be of the form $c(x,y) = -x^\top y$, the definition (1.11) of c-conjugate reduces to $\psi_0^c = -(-\phi_0)^*$, where we recall that for any convex function h on $\mathbb{R}^d$, $h^*(y) = \sup_{x\in\mathbb{R}^d}\{x^\top y - h(x)\}$ denotes its Legendre-Fenchel transform. It is well-known that if h is also lower semi-continuous and not identically infinite, the supremum in the definition of $h^*(y)$ is achieved by a point in its subdifferential; specifically, one has the relation

$$y \in \partial h(x) \iff x \in \partial h^*(y) \iff x^\top y = h(x) + h^*(y).$$

To derive analogous notions for c-concave functions, define the c-*superdifferential* of a c-concave function $f : \mathcal{X} \to \overline{\mathbb{R}}$ by

$$\partial^c f = \{(x,y) \in \mathcal{X} \times \mathcal{Y} : c(v,y) - f(v) \geq c(x,y) - f(x), \ \forall v \in \mathcal{X}\}.$$

Furthermore, let $\partial^c f(x) = \{y \in \mathbb{R}^d : (x,y) \in \partial^c f\}$ and $\partial^c f(B) = \bigcup_{x \in B} \partial^c f(x)$, for all $B \subseteq \mathcal{X}$. The following Lemma summarizes the main properties of c-concave functions which we shall require. Some of the statements which follow are weaker than necessary, but sufficient for our purposes.

Lemma 17. *Let $f : \mathcal{X} \to \overline{\mathbb{R}} := \mathbb{R} \cup \{-\infty\}$ be c-concave, and assume condition (**H0**).*

(i) (Villani (2008), Proposition 5.8) We have, $f^{cc} = f$.

(ii) (Villani (2003), Remark 1.13) Assume the cost c is bounded. Then, the supremum in the Kantorovich dual problem (1.9) *is achieved by a c-concave function $\phi \in L^1(P)$ such that $0 \leq \phi \leq \|c\|_\infty$ and $-\|c\|_\infty \leq \phi^c \leq 0$.*

(iii) (Gangbo and McCann (1996), Theorem 2.7) $\partial^c f$ is c-cyclically monotone, in the sense that for any permutation σ on $k \geq 1$ letters and any $(x_1,y_1), \dots, (x_k,y_k) \in \partial^c f$,

$$\sum_{j=1}^k c(x_j,y_j) \leq \sum_{j=1}^k c(x_{\sigma(j)},y_j).$$

(iv) (Gangbo and McCann (1996), Proposition C.4) Assume further that h is superlinear. For any given $x \in \mathbb{R}^d$, assume there exists a neighborhood of x over which f is bounded. Then, the c-superdifferential $\partial^c f(x)$ is nonempty. Therefore, if f is locally bounded over $\mathbb{R}^d$, it holds that for all $x, y \in \mathbb{R}^d$,

$$y \in \partial^c f(x) \iff x \in \partial^c f^c(y) \iff c(x,y) = f(x) + f^c(y).$$

In particular,

$$f(x) = \inf_{y \in \partial^c f(x)} \big\{c(x,y) - f^c(y)\big\}, \quad f^c(y) = \inf_{x \in \partial^c f^c(y)} \big\{c(x,y) - f(x)\big\}.$$

Furthermore, if f is in fact an optimal Kantorovich potential for the optimal transport problem from P to Q, and if $\mathcal{T}_c(P,Q) < \infty$, then for any optimal coupling $\pi \in \Pi(P,Q)$, $\mathrm{supp}(\pi) \subseteq \partial^c f$.

3.2 Upper Bounds for Compactly Supported Measures

We begin by bounding the rate of convergence of the empirical optimal transport cost in the special case where $\mathcal{X}$, $\mathcal{Y}$ and $\mathcal{Z} = \mathcal{X} - \mathcal{Y} = \{x - y : x \in \mathcal{X}, y \in \mathcal{Y}\}$ satisfy the following condition.

(**S1**) $\mathcal{X}, \mathcal{Y} \subseteq \mathbb{R}^d$ are convex and compact sets with nonempty interior. Furthermore, we have $\mathcal{X}, \mathcal{Y}, \mathcal{Z} \subseteq B_{0,1}$.

The assumption of compactness of $\mathcal{X}$ and $\mathcal{Y}$ will be relaxed in the following section, under concentration and anticoncentration conditions on the measures. Once $\mathcal{X}$ and $\mathcal{Y}$ are assumed compact, notice that the final assumption of (**S1**) can always be satisfied up to rescaling and recentering. Furthermore, we recall that the supports of P and Q are merely assumed to be contained in $\mathcal{X}$ and $\mathcal{Y}$, and thus need not be convex themselves.

We shall also assume throughout this section that the cost c satisfies condition (**H0**) and the following smoothness condition.

(**H1**) There exists $\alpha \in (0, 2]$ and a convex open set $\mathcal{Z}_1$ such that $\mathcal{Z} \subseteq \mathcal{Z}_1 \subseteq B_{0,2}$, and $h \in \mathscr{C}^\alpha(\mathcal{Z}_1)$. Furthermore, we have $0 \le h \le 1$ on $\mathcal{Z}_1$. We write $\Lambda := 1 \vee \|h\|_{\mathscr{C}^\alpha(\mathcal{Z}_1)} < \infty$.

For any measures $P \in \mathcal{P}(\mathcal{X})$ and $Q \in \mathcal{P}(\mathcal{Y})$, recall that $X_1, \dots, X_n \sim P$ and $Y_1, \dots, Y_n \sim Q$ denote i.i.d. samples, with corresponding empirical measures $P_n = \frac{1}{n}\sum_{i=1}^n \delta_{X_i}$ and $Q_n = \frac{1}{n}\sum_{i=1}^n \delta_{Y_i}$. The main result of this section is now stated as follows.

Theorem 9. *Assume conditions (**S1**), (**H0**), and (**H1**). Then, there exists a constant $C > 0$ depending only on $d, \alpha, \mathcal{X}, \mathcal{Y}, \mathcal{Z}_1$ such that*

$$\sup_{\substack{P \in \mathcal{P}(\mathcal{X}) \\ Q \in \mathcal{P}(\mathcal{Y})}} \mathbb{E}_{P,Q} \big| \mathcal{T}_c(P_n, Q_n) - \mathcal{T}_c(P, Q) \big| \le C \Lambda n^{-\alpha/d}.$$

Theorem 9 proves that the convergence rates anticipated in Section 3.1.1, for measures differing only in mean, in fact hold for all compactly supported measures. In particular, the $n^{-\alpha/d}$ rate of convergence is achievable as soon as $h \in \mathscr{C}^\alpha$, for $\alpha \in (0, 2]$, though is not generally claimed to improve further when $\alpha > 2$. For instance, the quadratic cost $\|\cdot\|^2$ lies in $\mathcal{C}^\infty$, but one cannot hope for a faster convergence rate than $n^{-2/d}$. Indeed, we derive matching lower bounds in Section 3.4 under closely related assumptions on the cost function c, which imply that the upper bound of Theorem 9 is generally unimprovable.

A careful investigation of our proof reveals that Theorem 9 in fact continues to hold for nonconvex costs h. We nevertheless prefer to retain the assumption of convexity in condition (**H0**) since it is required for the remainder of our main results; in particular, we do not claim that the convergence rate in Theorem 9 is sharp when h is not convex.

By letting α vanish, Theorem 9 suggests that the empirical optimal transport cost does not generally converge at any polynomial rate for cost functions which fail to be uniformly Hölder continuous. Indeed, absent any smoothness assumptions on c, $\mathcal{T}_c(P_n, Q_n)$ may not even converge in $L^1(\mathbb{P})$, as can be seen by taking c to be the Hamming metric. In this case, $\mathcal{T}_c(P_n, Q_n)$ is simply the Total Variation distance between P_n and Q_n, which almost surely equals unity when P and Q are absolutely continuous with respect to the Lebesgue measure.

As discussed in Section 3.1, perhaps the most widely-used cost functions satisfying conditions (**H0**) and (**H1**) are norms over $\mathbb{R}^d$ raised to a power greater than one. We illustrate the conclusion of Theorem 9 for such an example.

Corollary 3 (Powers of $\boldsymbol{\ell_r}$ Norms). *Let $\mathcal{X}, \mathcal{Y}$ satisfy condition (**S1**), and define the cost $c_{p,r}(x,y) = \|x-y\|_{\ell_r}^p$ for all $x \in \mathcal{X}, y \in \mathcal{Y}$ and $p, r \geq 1$. Let $P \in \mathcal{P}(\mathcal{X})$, $Q \in \mathcal{P}(\mathcal{Y})$.*

(i) We have for all $p, r \geq 1$, $\mathbb{E}\big|\mathcal{T}_{c_{p,r}}(P_n,Q_n) - \mathcal{T}_{c_{p,r}}(P,Q)\big| \lesssim n^{-(2\wedge p \wedge r)/d}$. In particular, specializing to $r=2$,

$$\mathbb{E}\big|W_p^p(P_n,Q_n) - W_p^p(P,Q)\big| \lesssim \begin{cases} n^{-p/d}, & 1 \leq p < 2 \\ n^{-2/d}, & 2 \leq p < \infty. \end{cases}$$

(ii) If $\mathcal{X}, \mathcal{Y} \subseteq \mathbb{R}^d$ are disjoint, then for all $p \geq 1$ and $r \geq 2$,

$$\mathbb{E}\big|\mathcal{T}_{c_{p,r}}(P_n,Q_n) - \mathcal{T}_{c_{p,r}}(P,Q)\big| \lesssim n^{-2/d}.$$

The proof is deferred to Section 3.A.1. Corollary 3(i) follows from the fact that $\|\cdot\|_{\ell_r}^p \in \mathscr{C}^{2\wedge p\wedge r}(\mathcal{Z}_1)$ for any bounded open set $\mathcal{Z}_1$, for all $p, r \geq 1$. When $r \geq 2$, notice that $\|\cdot\|_{\ell_r}$ is smooth away from the origin, so that the condition $\|\cdot\|_{\ell_r}^p \in \mathscr{C}^2(\mathcal{Z}_1)$ can be satisfied for any $p \geq 1$ whenever the closed set $\mathcal{Z} \subseteq \mathcal{Z}_1$ does not contain the point zero. This observation leads to Corollary 3(ii). This last point implies the rather surprising fact that for all measures P and Q admitting disjoint and compact support, one has

$$\mathbb{E}\big|W_1(P_n,Q_n) - W_1(P,Q)\big| \lesssim n^{-2/d}. \tag{3.14}$$

When the measures P and Q are not vanishingly close, Corollary 3 also translates into convergence rates for empirical Wasserstein distances.

Corollary 4 (Wasserstein Distances). *Let $p \geq 1$. Let $\mathcal{X}, \mathcal{Y}$ satisfy condition (**S1**), and let $P \in \mathcal{P}(\mathcal{X})$, $Q \in \mathcal{P}(\mathcal{Y})$. Assume $W_p(P,Q) \geq \delta_0$, for some constant $\delta_0 > 0$. Then,*

$$\mathbb{E}\big|W_p(P_n,Q_n) - W_p(P,Q)\big| \lesssim \delta_0^{1-p} \begin{cases} n^{-p/d}, & 1 \leq p < 2 \\ n^{-2/d}, & 2 \leq p < \infty. \end{cases} \tag{3.15}$$

Proof. By the numerical inequality $|x-y| \leq y^{1-p}|x^p - y^p|$ for all $x, y \geq 0$, $p \geq 1$, one has

$$\mathbb{E}|W_p(P_n,Q_n) - W_p(P,Q)| \leq \delta_0^{1-p}\mathbb{E}\big|W_p^p(P_n,Q_n) - W_p^p(P,Q)\big|.$$

The claim thus follows from Corollary 3. □

3.2.1 Proof of Theorem 9

We divide our argument into three cases.

3.2.1.1 Case 1: $\alpha = 2$

Under conditions (**S1**), (**H0**) and (**H1**), it follows from Lemma 17(ii) that there exist Kantorovich potentials $\phi_n : \mathcal{X} \to \mathbb{R}$ and $\psi_n : \mathcal{Y} \to \mathbb{R}$ such that $\mathcal{T}_c(P_n,Q_n) = J_{P_n,Q_n}(\phi_n,\psi_n)$ and

$|\phi_n|, |\psi_n| \leq 1$. Furthermore, since P and Q are compactly supported, it follows immediately from the definition of the c-conjugate that $(\phi_n, \psi_n) \in \Phi_c(P, Q)$, whence

$$\begin{aligned} \mathcal{T}_c(P,Q) &= \sup_{(\phi,\psi)\in\Phi_c(P,Q)} J_{P,Q}(\phi,\psi) \\ &\geq J_{P,Q}(\phi_n,\psi_n) = J_{P_n,Q_n}(\phi_n,\psi_n) + \int \phi_n d(P-P_n) + \int \psi_n d(Q-Q_n). \end{aligned} \tag{3.16}$$

On the other hand, recalling that $\mathcal{T}_c(P_n, Q_n) = J_{P_n,Q_n}(\phi_n, \psi_n)$, we derive

$$\mathcal{T}_c(P_n,Q_n) - \mathcal{T}_c(P,Q) \leq \int \phi_n d(P_n - P) + \int \psi_n d(Q_n - Q). \tag{3.17}$$

Our goal is now to bound the empirical processes arising on the right-hand side of the above display. Due to the compactness of $\mathcal{X}$ and $\mathcal{Y}$, it can be deduced from Gangbo and McCann (1996) that ϕ_n and ψ_n are Lipschitz and semi-concave. The following Lemma is an analogue of their results with explicit constants, whose proof is included in Section 3.A.2 for completeness.

Lemma 18. *Assume conditions (**S1**) and (**H0**), and that condition (**H1**) holds with $\alpha = 2$. Then the maps*

$$\widetilde{\phi}_n : x \in \mathcal{X} \longmapsto \phi_n(x) - \frac{\Lambda}{2}\|x\|^2, \quad \widetilde{\psi}_n : y \in \mathcal{Y} \longmapsto \psi_n(y) - \frac{\Lambda}{2}\|y\|^2$$

are concave and (2Λ)*-Lipschitz. Furthermore,* $|\widetilde{\phi}_n|, |\widetilde{\psi}_n| \leq 2\Lambda$.

For any $L, U > 0$, let $\mathcal{F}_{L,U}(K)$ denote the set of L-Lipschitz convex functions $f : K \to \mathbb{R}$ over a convex set $K \subseteq \mathbb{R}^d$, such that $|f| \leq U$. Recalling the convexity of $\mathcal{X}$ and $\mathcal{Y}$ under condition (**S1**), define

$$\Delta_n = \sup_{f\in\mathcal{F}_{1,1}(\mathcal{X})} \int f d(P_n - P) + \sup_{g\in\mathcal{F}_{1,1}(\mathcal{Y})} \int g d(Q_n - Q).$$

By Lemma 18, we have $(-\widetilde{\phi}_n/2\Lambda) \in \mathcal{F}_{1,1}(\mathcal{X})$ and $(-\widetilde{\psi}_n/2\Lambda) \in \mathcal{F}_{1,1}(\mathcal{Y})$, thus together with equation (3.17) we obtain

$$\mathcal{T}_c(P_n,Q_n) - \mathcal{T}_c(P,Q) \leq 2\Lambda\Delta_n + \frac{\Lambda}{2}\int \|\cdot\|^2 d\big((P_n - P) + (Q_n - Q)\big). \tag{3.18}$$

On the other hand, lower bounds on $\mathcal{T}_c(P_n,Q_n) - \mathcal{T}_c(P,Q)$ are simple to obtain. As before, there exists a pair of optimal Kantorovich potentials $(\phi_0, \psi_0) \in \Phi_c(P,Q)$ such that $|\phi_0| \vee |\psi_0| \leq 1$ and $\mathcal{T}_c(P,Q) = J_{P,Q}(\phi_0,\psi_0)$. Therefore,

$$\mathcal{T}_c(P_n,Q_n) \geq J_{P_n,Q_n}(\phi_0,\psi_0) = \mathcal{T}_c(P,Q) + \int \phi_0 d(P_n - P) + \int \psi_0 d(Q_n - Q).$$

Combining the previous two displays, we deduce

$$\mathbb{E}\big|\mathcal{T}_c(P_n,Q_n) - \mathcal{T}_c(P,Q)\big| \leq 2\Lambda\mathbb{E}[\Delta_n] + \frac{\Lambda}{2}\mathbb{E}\left|\int \|\cdot\|^2 d\big((P_n - P) + (Q_n - Q)\big)\right|$$

$$+\mathbb{E}\left|\int \phi_0 d(P_n - P)\right| + \mathbb{E}\left|\int \psi_0 d(Q_n - Q)\right| \lesssim \Lambda\Big(\mathbb{E}[\Delta_n] + n^{-1/2}\Big), \tag{3.19}$$

where the final bound is a straightforward consequence of Chebyshev's inequality, and where we have used the fact that $|\phi_0|, |\psi_0| \leq 1$. It thus remains to bound $\mathbb{E}[\Delta_n]$. This last is a sum of expected suprema of empirical processes indexed by convex Lipschitz functions, upper bounds for which can be obtained via Dudley's chaining technique (Dudley, 2014) in terms of the metric entropy of the class $\mathcal{F}_{L,U}(K)$. Specifically, recall that for all $\epsilon > 0$, the ϵ-metric entropy of a set A contained in a metric space $(\mathcal{X}, \eta)$ is the logarithm of the ϵ-covering number $N(\epsilon, A, \eta)$ of (A, η), defined by

$$N(\epsilon, A, \eta) = \inf\big\{N \geq 1 : \exists\{x_1, \dots, x_N\} \subseteq \mathcal{X}, \forall x \in A, \exists 1 \leq i \leq N : \eta(x, x_i) \leq \epsilon\big\}.$$

The following version of Dudley's bound will be sufficient for our purposes, and can be deduced for instance from Lemma 16 of von Luxburg and Bousquet (2004) (see also Lemma 3.2 of van de Geer (2000)).

Lemma 19 (von Luxburg and Bousquet (2004))**.** *Let $\mathcal{G}$ be a set of real-valued measurable functions on $\mathbb{R}^d$. Then,*

$$\mathbb{E}\left[\sup_{g\in\mathcal{G}} \int g d(P_n - P)\right] \lesssim \mathbb{E}\left[\inf_{\tau>0}\left\{\tau + \frac{1}{\sqrt{n}}\int_\tau^\infty \sqrt{\log N(\epsilon, \mathcal{G}, L^2(P_n))}d\epsilon\right\}\right]. \tag{3.20}$$

Tight bounds on the metric entropy of the class $\mathcal{F}_{L,U}(K)$ are well known, and were first obtained in general dimension d by Bronshtein (1976) (see also Dudley (2014)). The following is a version of Bronshtein's result stated in Theorem 1 of Guntuboyina and Sen (2012) with explicit dependence on the constants L and U, which we shall also use in Section 3.3.

Lemma 20 (Bronshtein (1976))**.** *There exist universal constants $C, \epsilon_0 > 0$ such that for every $L, U > 0$ and $b > a$, we have for all $\epsilon \leq \epsilon_0(U + L(b-a))$,*

$$\log N\big(\epsilon, \mathcal{F}_{L,U}([a,b]^d), L^\infty\big) \leq C\left(\frac{U + L(b-a)}{\epsilon}\right)^{\frac{d}{2}}.$$

Notice that, by condition (**S1**),

$$N(\cdot, \mathcal{F}_{1,1}(\mathcal{X}), L^2(P_n)) \leq N(\cdot, \mathcal{F}_{1,1}(\mathcal{X}), L^\infty) \leq N(\cdot, \mathcal{F}_{1,1}([-1,1]^d), L^\infty),$$

and the same upper bound holds for the covering number $N(\cdot, \mathcal{F}_{1,1}(\mathcal{Y}), L^2(Q_n))$. Combine these facts with equation (3.19) and with Lemmas 19–20 to deduce that for any $\tau > 0$,

$$\mathbb{E}\big|\mathcal{T}_c(P_n, Q_n) - \mathcal{T}_c(P, Q)\big| \lesssim \Lambda\left[n^{-1/2} + \tau + \frac{1}{\sqrt{n}}\int_\tau^\infty \epsilon^{-\frac{d}{4}} d\epsilon\right] \lesssim \Lambda\left[n^{-1/2} + \tau + \frac{\tau^{1-\frac{d}{4}}}{\sqrt{n}}\right], \tag{3.21}$$

where we have used the assumption $d \geq 5$. Choosing $\tau \asymp n^{-2/d}$ leads to the claimed bound,

$$\mathbb{E}\big|\mathcal{T}_c(P_n, Q_n) - \mathcal{T}_c(P, Q)\big| \lesssim \Lambda n^{-2/d}.$$

3.2.1.2 Case 2: $1 < \alpha < 2$

We prove the claim using a smooth approximation of the cost h, thereby appealing to the result of Case 1. Let $K : \mathbb{R}^d \to \mathbb{R}_+$ be an even, smooth mollifier with support lying in $B_{0,1}$. For $\sigma > 0$, write $K_\sigma(x) = \sigma^{-d}K(x/\sigma)$. Define the modified cost function $c_\sigma(x, y) = h_\sigma(x - y)$, where $h_\sigma = h \star K_\sigma$.

Lemma 21. *Assume conditions **(S1)**, **(H0)**–**(H1)** hold for some $\alpha \in (1, 2)$. Then, there exist universal constants $C > 0$ and $\epsilon \in (0, 1)$ such that for all $\sigma \in (0, \epsilon)$, the following statements hold.*

(i) We have, $\|h - h_\sigma\|_{L^\infty(\mathcal{Z})} \le \Lambda\sigma^\alpha$.

*(ii) The cost function c_σ itself satisfies condition **(H0)**, and satisfies condition **(H1)** in the sense that there exists an open set $\widetilde{\mathcal{Z}}_1 \supseteq \mathcal{Z}$ contained in $B_{0,2}$ such that $h_\sigma \le 1$ on $\widetilde{\mathcal{Z}}_1$ and*

$$\|h_\sigma\|_{\mathscr{C}^2(\widetilde{\mathcal{Z}}_1)} \le \Lambda_\sigma := C\Lambda\sigma^{\alpha-2}.$$

The proof of Lemma 21 appears in Section 3.A.3. Notice that

$$\sup_{\substack{\widetilde{P}\in\mathcal{P}(\mathcal{X})\\ \widetilde{Q}\in\mathcal{P}(\mathcal{Y})}} |\mathcal{T}_c(\widetilde{P}, \widetilde{Q}) - \mathcal{T}_{c_\sigma}(\widetilde{P}, \widetilde{Q})| \le \|h - h_\sigma\|_{L^\infty(\mathcal{Z})},$$

thus Lemma 21(i) implies

$$\begin{aligned}
&\mathbb{E}\big|\mathcal{T}_c(P_n, Q_n) - \mathcal{T}_c(P, Q)\big| \\
&\quad\le \mathbb{E}\big|\mathcal{T}_c(P_n, Q_n) - \mathcal{T}_{c_\sigma}(P_n, Q_n)\big| + \mathbb{E}\big|\mathcal{T}_{c_\sigma}(P_n, Q_n) - \mathcal{T}_{c_\sigma}(P, Q)\big| + \mathbb{E}\big|\mathcal{T}_{c_\sigma}(P, Q) - \mathcal{T}_c(P, Q)\big| \\
&\quad\le 2\Lambda\sigma^\alpha + \mathbb{E}\big|\mathcal{T}_{c_\sigma}(P_n, Q_n) - \mathcal{T}_{c_\sigma}(P, Q)\big|.
\end{aligned}$$

On the other hand, by Lemma 21(ii), we may apply the result of Case 1 to obtain,

$$\mathbb{E}\big|\mathcal{T}_{c_\sigma}(P_n, Q_n) - \mathcal{T}_{c_\sigma}(P, Q)\big| \lesssim \Lambda\sigma^{\alpha-2}n^{-2/d}.$$

Altogether, we deduce that for any $\sigma \in (0, \epsilon)$,

$$\mathbb{E}\big|\mathcal{T}_c(P_n, Q_n) - \mathcal{T}_c(P, Q)\big| \lesssim \Lambda\left[\sigma^\alpha + \sigma^{\alpha-2}n^{-\frac{2}{d}}\right].$$

Choosing $\sigma \asymp n^{-1/d}$ leads to an upper bound scaling at the rate $\Lambda n^{-\alpha/d}$ on the right-hand side of the above display, as claimed.

3.2.1.3 Case 3: $0 < \alpha \le 1$

When $\alpha \in (0, 1]$, the claim follows by a simpler argument than that of Case 1. For any bounded set $K \subseteq \mathbb{R}^d$ and $L > 0$, define the α-Hölder ball $\mathcal{C}^\alpha(K; L) = \{f \in \mathcal{C}^\alpha(K) : \|f\|_{\mathcal{C}^\alpha(K)} \le L\}$, and let ϕ_n and ψ_n be defined as in Case 1. When $\alpha < 2$, Lemma 18 no longer guarantees that these potentials are semi-concave, however the following Hölder estimate is easily derived, and stated without proof.

Lemma 22. *Assume conditions (**S1**) and (**H0**), and that condition (**H1**) holds with $\alpha \in (0,1]$. Then, $\phi_n \in \mathcal{C}^\alpha(\mathcal{X};\Lambda)$ and $\psi_n \in \mathcal{C}^\alpha(\mathcal{Y};\Lambda)$ for all $n \geq 1$.*

By an analogous reduction as in Case 1, we therefore have for any $\tau > 0$,

$$\mathbb{E}\big|\mathcal{T}_c(P_n,Q_n) - \mathcal{T}_c(P,Q)\big| \lesssim \Lambda\left[n^{-1/2} + \tau + \frac{1}{\sqrt{n}}\int_\tau^\infty \sqrt{\log N(\epsilon, \mathcal{C}^\alpha([-1,1]^d;1), L^\infty)}d\epsilon\right].$$

By Theorem 2.7.1 of van der Vaart and Wellner (1996), one has

$$\log N(\epsilon, \mathcal{C}^\alpha([-1,1]^d;1), L^\infty) \lesssim \epsilon^{-d/\alpha}, \quad \epsilon > 0,$$

implying that the right-hand side of the penultimate display is of order $\Lambda n^{-\alpha/d}$ if $\tau \asymp n^{-\alpha/d}$ and $d \geq 5 > 2\alpha$. The claim follows. □

Remark 1. Though Theorem 11 is only stated when $d \geq 5$, a simple extension of our proof yields the rate $n^{-1/2}$ whenever $d \leq 4$ and $\alpha = 2$, up to a logarithmic factor when $d = 4$; this follows by taking $\tau = 0$ in equation (3.21). A similar extension can be made when $\alpha \in (0,1]$. On the other hand, our mollification step for the case $\alpha \in (1,2)$ does not appear to yield a sharp convergence rate when $d < 2\alpha < 5$. After a preprint of this chapter was made publicly available, the work of Hundrieser, Staudt, and Munk (2022) has extended Theorem 9, by showing that in all dimensions $d \geq 1$, and for all $\alpha \in (0,2]$, the upper bound

$$\mathbb{E}\big|\mathcal{T}_c(P_n,Q_n) - \mathcal{T}_c(P,Q)\big| \lesssim \begin{cases} n^{-\alpha/d}, & 2\alpha < d \\ n^{-\alpha/d}\log n, & 2\alpha = d \\ n^{-1/2}, & 2\alpha > d \end{cases} \tag{3.22}$$

holds under the same assumptions as those of Theorem 9. In particular, this result recovers our Theorem 9 when $d \geq 5$.

3.3 Upper Bounds for Unbounded Measures under Tail Conditions

We now turn to upper bounding the rate of convergence of the empirical optimal transport cost for measures $P, Q \in \mathcal{P}(\mathbb{R}^d)$ with unbounded support, under suitable tail conditions. We shall assume that the cost function c satisfies the following smoothness assumption, which is a suitable generalization of condition (**H1**) to the present setting.

(**H2**) $h \in \mathscr{C}^2_{\text{loc}}(\mathbb{R}^d)$, and there exist $p, \Lambda \geq 1$ such that $\|h\|_{\mathscr{C}^2(B_{0,r})} \leq \Lambda r^p$ for all $r \geq 1$.

Notice that unlike in Section 3.2, we limit our exposition to costs lying in $\mathscr{C}^2_{\text{loc}}$ rather than $\mathscr{C}^\alpha_{\text{loc}}$ for all $\alpha \in (0,2]$. As we shall see, our main result is nevertheless sufficiently general to cover the costs $h(x) = \|x\|^p$ for all $p > 1$, via an approximation argument.

Our upper bounds in Section 3.2 hinged upon Lemma 18, which provided quantitative estimates on the Lipschitz and semi-concavity constants of optimal Kantorovich potentials, for any sufficiently smooth cost function over a compact set. In contrast, we now only assume a local Hölder estimate on c in assumption (**H2**), thus the optimal Kantorovich potentials between two measures on $\mathbb{R}^d$ will generally not be *globally* Lipschitz or semi-concave. While these properties nevertheless hold *locally* under rather general conditions (Gangbo and McCann, 1996), we are not aware of existing quantitative estimates under the conditions required for our development. One of our key technical contributions in this section is to obtain such quantitative bounds, which we now describe before stating our main result. We begin with the following straightforward generalization of Lemma 18, whose proof appears in Section 3.B.1.

Lemma 23. *Let c be a cost function satisfying conditions (**H0**) and (**H2**) with h superlinear. Given two measures $P, Q \in \mathcal{P}(\mathbb{R}^d)$, let $\pi \in \Pi(P, Q)$ be an optimal coupling with respect to c, and assume there exists a locally bounded c-concave function $\phi : \mathbb{R}^d \to \mathbb{R}$ such that $\mathrm{supp}(\pi) \subseteq \partial^c\phi$. Let $r \geq 1$, and let*

$$\Lambda_r = \Lambda \sup\left\{\|x - y\|^p : x \in B_{0,r},\ y \in \partial^c\phi(B_{0,r})\right\}.$$

Then, there exist universal constants $c_1, c_2 > 0$ such that the map

$$x \in B_{0,r} \mapsto \phi(x) - c_1\Lambda_r \|x\|^2,$$

is concave, and Lipschitz with parameter $c_2 r\Lambda_r$.

Lemma 23 shows that the local Lipschitz and semi-concavity constants for a Kantorovich potential ϕ are largely driven by the maximal displacement induced by the coupling π over points lying in $B_{0,r}$. We will show how L^∞ estimates on these displacements can be obtained under the following conditions on P, Q.

(i) **Super-Gaussian Anticoncentration.** We will say a measure P is (γ, b)-super-Gaussian for some $\gamma, b > 0$ if for any $x \in \mathbb{R}^d$,

$$P(B_x) \geq b \cdot \mathbb{P}(Z \in B_x), \quad \text{where } Z \sim N(0, \gamma^2).$$

(ii) **Sub-Weibull Concentration.** A measure Q is said to be (σ, β)-sub-Weibull (Kuchibhotla and Chakrabortty, 2022; Vladimirova et al., 2020) for some $\sigma > 0$ and $0 < \beta \leq 2$ if

$$\int \exp\left[\frac{1}{2}\left(\frac{\|y\|}{\sigma}\right)^\beta\right] dQ(y) \leq 2.$$

The assumption of super-Gaussianity implies an *anticoncentration* bound for the underlying measure, in the sense that it cannot place significantly less probability mass than a Gaussian distribution in any unit-radius ball. In Section 3.B.9, we show that a measure is super-Gaussian whenever it admits a regular Lebesgue density in the sense of Polyanskiy and Wu (2016). For instance, Polyanskiy and Wu show that for any probability measure P with finite first moment, the mixture distribution $K_\tau \star P$ admits a regular density, where K_τ is the $N(0, \tau^2 I_d)$ density for some $\tau > 0$. Any such measure is thus also super-Gaussian. We also note that absolute continuity is not necessary for super-Gaussianity; for example, given any (possibly atomic) measure $\rho \in \mathcal{P}(\mathbb{R}^d)$, any super-Gaussian measure P, and any $\lambda \in (0,1)$, the measure $\lambda\rho + (1-\lambda)P$ is also super-Gaussian.

On the other hand, the sub-Weibull condition is a *concentration* assumption which generalizes the well-known sub-Gaussian and sub-exponential conditions, which respectively correspond to the cases $\beta = 2$ and $\beta = 1$ up to rescaling of the constant σ (Boucheron, Lugosi, and Massart, 2013). Indeed, it is a straightforward consequence of Markov's inequality that if Y has (σ, β)-sub-Weibull distribution for some $\beta \in (0, 2]$ and $\sigma > 0$, then for all $u > 0$,

$$\mathbb{P}(\|Y\| \geq u) \leq 2 \exp\left\{ -\frac{1}{2}\left(\frac{u}{\sigma}\right)^{\beta} \right\}. \tag{3.23}$$

Finally, we shall require the following condition on the cost function c.

(H3) We have $h(0) = 0$. Furthermore, there exist constants $p > 1$, $\kappa \geq 1$, and a convex differentiable function $\omega : (1, \infty) \to (1, \infty)$ such that

$$h(z) = \omega(\|z\|), \quad \text{and} \quad \frac{1}{\kappa}\|z\|^{p-1} \leq \omega'(\|z\|) \leq \kappa \|z\|^{p-1} \quad \text{for all } \|z\| > 1.$$

Condition (**H3**) implies that the function h is superlinear, with order of growth comparable to that of $\|\cdot\|^p$ for some $p > 1$. Aside from the assumption $h(0) = 0$, which can always be satisfied by translation, we emphasize that condition (**H3**) does not constrain the behaviour of h near zero, but is nevertheless stronger than the conditions assumed in Section 3.2. Therefore, we provide several examples of cost functions satisfying these two conditions before turning to our main results.

The most important example of a cost satisfying our assumptions is, of course, $\|\cdot\|^p$ for $p \geq 2$. However, when $p \in (1, 2)$, the cost $\|\cdot\|^p$ does not satisfy condition (**H2**); to study this case, we will employ an approximation argument with the cost function $h_\epsilon(x) = (\|x\|^2 + \epsilon^{2/p})^{p/2} - \epsilon$, which satisfies (**H2**)–(**H3**) for any $\epsilon > 0$ and $p > 1$. This cost function has been of interest in its own right in the optimal transport literature, as it forms an approximation of $\|\cdot\|^p$ which satisfies the celebrated Ma-Trudinger-Wang regularity conditions even when $p \neq 2$ (Ma, Trudinger, and Wang, 2005; Li, Santambrogio, and Wang, 2014).

Conditions (**H2**)–(**H3**) also hold for costs that have different power-type behaviors at the origin and infinity, such as $h(z) = \lambda_p \|z\|^p + \lambda_q \|z\|^q$ for $p > q \geq 2$, which arise in the study of modified transport problems with congestion costs (Brasco, Carlier, and Santambrogio, 2010; Carlier, Jimenez, and Santambrogio, 2008).

More generally, conditions (**H2**)–(**H3**) are satisfied by any twice continuously differentiable convex cost function of the form

$$h(x) \propto \begin{cases} \|x\|^p, & \|x\| > 1 \\ h_0(x), & \|x\| \leq 1, \end{cases}$$

where $h_0(0) = 0$ and $p \geq 2$. This family includes, for instance, smooth approximations of the truncated cost $h(x) = \|x\|^p I(x \in B_0^c)$, and ℓ_p analogues of Huber's loss function.

While the smoothness condition (**H2**) will be needed in order to appeal to Lemma 23, assumption (**H3**) is sufficient to obtain the following result, which plays a central role in our development.

Theorem 10. *Let $P, Q \in \mathcal{P}(\mathbb{R}^d)$ and assume Q is (σ, β)-sub-Weibull. Let c be a cost function satisfying conditions (**H0**) and (**H3**), and let $\pi \in \Pi(P, Q)$ have c-cyclically monotone support. Then, there exists a constant $C > 0$ depending on d, p, β, κ such that for any c-concave function ϕ satisfying* $\mathrm{supp}(\pi) \subseteq \partial^c\phi$*,*

$$\sup_{y \in \partial^c\phi(x)} \|y\| \leq C\sigma \left\{ (\|x\| + 1) \vee \sup_{\|x-y\| \leq 2} \left[\log\left(\frac{1}{P(B_y)} \right) \right]^{\frac{1}{\beta}} \right\}, \quad x \in \mathbb{R}^d. \tag{3.24}$$

In particular, if P is (γ, b)-super-Gaussian, h is strictly convex, and $\mathcal{T}_c(P, Q) < \infty$, then the unique optimal transport map T pushing P forward onto Q satisfies for P-a.e. $x \in \mathbb{R}^d$,

$$\|T(x)\| \leq C'\sigma(\|x\| + 1)^{\frac{2}{\beta}}, \tag{3.25}$$

for a constant $C' > 0$ depending on $d, \kappa, p, \beta, \gamma, b$.

We defer the proof to Section 3.3.2. Theorem 10 implies that any optimal transport plan between P and Q does not move probability mass from any point $x \in \mathbb{R}^d$ by more than a polynomial of $\|x\|$. To obtain this result, we required an anticoncentration assumption on the source measure P and a concentration assumption on the target measure Q, ensuring that their tails are sufficiently comparable to avoid large transports of mass. It is easy to see that assumptions of this nature are necessary: for instance, if P were compactly supported and Q were supported over $\mathbb{R}^d$, any transport plan in $\Pi(P, Q)$ would couple a nonzero amount of mass from the bounded support of P with points lying at an arbitrarily far distance.

In the special case where Q is sub-Gaussian, its tails are no heavier than those of a super-Gaussian measure P. Equation (3.25) shows that the optimal transport map from P to Q grows at most linearly in this regime, irrespective of the order of growth p of the cost function. This bound is clearly unimprovable in general, as can be seen by taking $P = Q$.

Our proof of Theorem 10 is inspired by its non-quantitative analogues proven by Gangbo and McCann (1996), and by Colombo and Fathi (2021) who derived analogous quantitative bounds for the special case where P is a Gaussian measure and $h(x) = \|x\|^2$. Unlike Colombo and Fathi (2021), our result holds for any cost function satisfying conditions (**H0**) and (**H3**), and for general measures P, Q which are not presumed to be absolutely continuous with respect to the Lebesgue measure. We shall require this level of generality in the sequel, when Theorem 10 will be invoked for P and Q replaced by their empirical counterparts.

Equipped with Theorem 10, we are ready to state the main result of this section.

Theorem 11. *Assume conditions (**H0**), (**H2**) and (**H3**) hold. Assume further that $P, Q \in \mathcal{P}(\mathbb{R}^d)$ are both (σ, β)-sub-Weibull, and (γ, b)-super-Gaussian. Then, there exists a constant $C > 0$ depending on $\sigma, \beta, \gamma, b, \kappa, p, d$ such that*

$$\mathbb{E}\big|\mathcal{T}_c(P_n, Q_n) - \mathcal{T}_c(P, Q)\big| \leq C\Lambda n^{-\frac{2}{d}}.$$

Theorem 11 shows that, for $\mathscr{C}^2_{\mathrm{loc}}$ convex costs with polynomial rate of growth, the $n^{-2/d}$ rate of convergence obtained for compactly supported measures in Section 3.2 carries over to unboundedly supported measures, with tails satisfying suitable concentration and anticoncentration conditions. While Theorem 11 does not provide upper bounds for $\mathscr{C}^\alpha_{\mathrm{loc}}$ costs when $\alpha < 2$, it is sufficiently general to deduce the following special case.

Corollary 5. *Assume $P, Q \in \mathcal{P}(\mathbb{R}^d)$ satisfy the same conditions as Theorem 11. Then, for all $p > 1$,*

$$\mathbb{E}\big|W_p^p(P_n, Q_n) - W_p^p(P, Q)\big| \lesssim \begin{cases} n^{-p/d}, & 1 < p \leq 2 \\ n^{-2/d}, & 2 \leq p < \infty. \end{cases}$$

For the regime $p \geq 2$, this result follows as a direct consequence of Theorem 11, while for the regime $1 < p < 2$, we achieve the claim by using a smooth uniform approximation of $\|\cdot\|^p$ which satisfies the conditions of Theorem 11. The proof is deferred to Section 3.B.7.

By reasoning identically as in Corollary 4, one can also deduce a convergence rate for empirical Wasserstein distances between measures with unbounded support. In particular, equation (3.15) continues to hold for all $P, Q \in \mathcal{P}(\mathbb{R}^d)$ satisfying the conditions of Theorem 11, and satisfying $W_p(P, Q) \geq \delta_0 > 0$.

3.3.1 Proof of Theorem 11

As in the proof of Theorem 9, we shall reduce the problem of bounding the $L^1(\mathbb{P})$ convergence rate of $\mathcal{T}_c(P_n, Q_n)$ to that of bounding the supremum of an empirical process. Unlike in Theorem 9, the relevant empirical process in this section will be indexed by *locally* semi-concave Lipschitz functions, with quantitative local Lipschitz and semi-concavity moduli obtained in part by appealing to Theorem 10. Our proof proceeds with eight steps, the first five of which carry out this reduction, and the last three of which bound the resulting empirical process. Throughout the proof, $C, C', C_i, c_i > 0, i \geq 0$, denote constants possibly depending on $\sigma, \beta, \gamma, b, \kappa, p, d$, which do not depend on Λ or otherwise on c, P and Q, and whose value may change from line to line. Likewise, the symbols $\lesssim$ and $\asymp$ hide constants possibly depending on the former quantities. All intermediary results appearing in the sequel are proven in Section 3.B.

Step 0: Setup. Let $L_j = [-3^j, 3^j]^d$ for all $j \geq 0$. For all $j \geq 1$, let $I_{j1}, \dots, I_{jm_d}$ denote the $m_d := 3^d - 1$ cubes of side-length $2 \cdot 3^{j-1}$ forming the natural partition of $L_j \setminus L_{j-1}$. For notational convenience, set $I_0 \equiv I_{0k} = L_0$ for all $k = 1, \dots, m_d$, and define

$$I_j := L_j \setminus L_{j-1} = \bigcup_{k=1}^{m_d} I_{jk}, \quad j \geq 1.$$

Note that $\{I_j : j \geq 0\}$ and $\{I_{jk} : j \geq 0, 1 \leq k \leq m_d\}$ are partitions of $\mathbb{R}^d$, up to measure-zero intersections. We also write $\ell_j = \sup_{x \in I_j} \|x\| = \sqrt{d}3^j$ for all $j \geq 0$.

Let $\mathcal{F}$ denote the set of convex functions over $\mathbb{R}^d$. Recall that $\mathcal{F}_{m,u}(I)$ denotes the set of m-Lipschitz convex functions over a convex set $I \subseteq \mathbb{R}^d$, which are uniformly bounded over I

by $u > 0$. Let $M = (M_j : j \geq 0)$ and $U = (U_j : j \geq 0)$ denote sequences of nonnegative real numbers, and let

$$\mathcal{K}_{M,U} = \Big\{ f : \mathbb{R}^d \to \mathbb{R} : (-f)|_{I_{jk}} \in \mathcal{F}_{M_j,U_j}(I_{jk}),\ j \geq 0, 1 \leq k \leq m_d \Big\}. \tag{3.26}$$

Step 1: Extension of Kantorovich Potentials. Let

$$R_n = \max_{1 \leq i \leq n} \|X_i\| \vee \|Y_i\|.$$

Conditions (**H0**) and (**H3**) imply that $h(z) \leq \kappa \|z\|^p$ for all $z \in \mathbb{R}^d$, thus h is bounded above by $\overline{R}_n := \kappa(2R_n)^p$ over $B_{0,2R_n}$. It can then be deduced from Lemma 17(ii) that there exist potentials $f_n : \mathrm{supp}(P_n) \to [-\overline{R}_n, 0]$ and $g_n : \mathrm{supp}(Q_n) \to [0, \overline{R}_n]$ such that $(f_n, g_n) \in \Phi_c(P_n, Q_n)$ and (f_n, g_n) is optimal for the optimal transport problem from P_n to Q_n. We extend the domain of f_n and g_n to $\mathbb{R}^d$ using the following construction. Define for all $y \in \mathbb{R}^d$,

$$\eta_n(y) = \inf_{x \in \mathrm{supp}(P_n)} \Big\{ c(x,y) - f_n(x) \Big\} \wedge \overline{R}_n,$$

and for all $x, y \in \mathbb{R}^d$,

$$\phi_n(x) = \eta_n^c(x) = \inf_{y \in \mathbb{R}^d} \Big\{ c(x,y) - \eta_n(y) \Big\}, \quad \psi_n(y) = \eta_n^{cc}(y) = \inf_{x \in \mathbb{R}^d} \Big\{ c(x,y) - \eta_n^c(x) \Big\}.$$

Lemma 24. *Given an optimal coupling π_n between P_n and Q_n, the following hold.*

(i) For all $x, y \in \mathbb{R}^d$, $\phi_n(x) + \psi_n(y) \leq c(x,y)$.

(ii) $\phi_n(x) = f_n(x)$ for all $x \in \mathrm{supp}(P_n)$, and $\psi_n(y) = g_n(y)$ for all $y \in \mathrm{supp}(Q_n)$. In particular,

$$\mathcal{T}_c(P_n, Q_n) = \int \phi_n dP_n + \int \psi_n dQ_n.$$

(iii) For all $x \in \mathbb{R}^d$, $|\phi_n(x)| \vee |\psi_n(x)| \leq \overline{R}_n$.

(iv) For all $(x,y) \in \mathrm{supp}(\pi_n)$, $(x,y) \in \partial^c \phi_n$ and $(y,x) \in \partial^c \psi_n$.

By Lemma 24(iii), ϕ_n and ψ_n are bounded, so $\phi_n \in L^1(P)$ and $\psi_n \in L^1(Q)$. This fact combined with Lemma 24(i) guarantees that $(\phi_n, \psi_n) \in \Phi_c(P,Q)$, whence

$$\begin{aligned}\mathcal{T}(P,Q) &= \sup_{(\phi,\psi) \in \Phi_c(P,Q)} \int \phi dP + \int \psi dQ \\ &\geq \int \phi_n dP + \int \psi_n dQ = \mathcal{T}_c(P_n, Q_n) + \int \phi_n d(P - P_n) + \int \psi_n d(Q - Q_n).\end{aligned} \tag{3.27}$$

It remains to bound the last two terms on the right-hand side of the above display. We shall do so by first proving that $\phi_n, \psi_n \in \mathcal{K}_{M,U}$ with high probability, for suitable sequences M and U. We focus on ϕ_n, and a symmetric argument can be used for ψ_n.

By Lemma 23, recall that the Lipschitz and semi-concavity moduli of $\phi_n|_{I_{jk}}$ are largely driven by the magnitude of the c-superdifferential $\partial^c\phi_n(I_{jk})$. The bulk of our effort will go into bounding this quantity. In fact, it will suffice to bound that of $\partial^c\phi_n(L_j)$, for all $j \geq 0$. To do so, we proceed with the following step, in view of invoking Theorem 10 with the measures P_n and Q_n.

Step 2: Global Concentration and Local Anticoncentration of $\boldsymbol{P_n}$, $\boldsymbol{Q_n}$. Fix $\rho = \frac{2p}{\beta} \vee \frac{d}{4}$, and set

$$V_{1,n} = \int \exp\left(\frac{\|x\|^\beta}{2\rho\sigma^\beta}\right) dP_n(x), \quad V_{2,n} = \int \exp\left(\frac{\|y\|^\beta}{2\rho\sigma^\beta}\right) dQ_n(y).$$

By Jensen's inequality, notice that

$$\int \exp\left(\frac{\|x\|^\beta}{2\rho V_{1,n}\sigma^\beta}\right) dP_n(x) \leq V_{1,n}^{1/V_{1,n}} \leq 2, \tag{3.28}$$

implying that P_n is $(\sigma(\rho V_{1,n})^{1/\beta}, \beta)$-sub-Weibull. Similarly, Q_n is $(\sigma(\rho V_{2,n})^{1/\beta}, \beta)$-sub-Weibull, implying that both are $(\sigma(\rho V_n)^{1/\beta}, \beta)$-sub-Weibull when $V_n = V_{1,n} + V_{2,n}$.

We further show that P_n satisfies a high-probability anticoncentration bound in a sufficiently small region about the origin. Specifically, define the integer

$$J_n = \left\lfloor \frac{1}{2}\log_3\left(\frac{\gamma^2}{4d}\log n\right)\right\rfloor, \tag{3.29}$$

and the event

$$A_n = \bigcap_{j=0}^{J_n} \left\{ \inf_{x\in L_j} \inf_{\|x-y\|\leq 2} P_n(B_y) \geq \frac{C_1}{2}\exp(-\ell_j^2/\gamma^2)\right\},$$

where we recall that the parameter γ arises from the super-Gaussianity assumption on P and Q. The following result is proven using elementary tools from empirical process theory.

Lemma 25. *There exists a sufficiently large choice of the constant $C_1 > 0$, depending only on γ, such that $\mathbb{P}(A_n^c) \lesssim 1/n$.*

Step 3: Bounding $\boldsymbol{\partial^c\phi_n}$. Step 2 will allow us to bound $\partial^c\phi_n(L_j)$ whenever $0 \leq j \leq J_n$ by invoking Theorem 10. On the other hand, P_n may place insufficient mass outside the box L_{J_n} to appeal to Theorem 10 when $j > J_n$, thus we treat this case separately below.

- **Regime 1: $\boldsymbol{0 \leq j \leq J_n}$.** By Step 2, Q_n is $(\sigma(\rho V_n)^{1/\beta}, \beta)$-sub-Weibull, and $\mathrm{supp}(\pi_n) \subseteq \partial^c\phi_n$ by Lemma 24(iv). Therefore, by Theorem 10, we have for all $0 \leq j \leq J_n$,

  $$\sup_{y\in\partial^c\varphi_n(L_j)} \|y\| \lesssim V_n^{\frac{1}{\beta}} \left\{ (\ell_j + 1) \vee \sup_{\|x-y\|\leq 2} \left[\log\left(\frac{1}{P_n(B_y)}\right)\right]^{\frac{1}{\beta}} \right\}.$$

  Over the event A_n, we therefore have uniformly in $0 \leq j \leq J_n$,

  $$\sup_{y\in\partial^c\phi_n(L_j)} \|y\| \lesssim V_n^{\frac{1}{\beta}} \left[\ell_j + 1 + \ell_j^2/\gamma^2\right]^{\frac{1}{\beta\wedge 1}} \lesssim V_n^{\frac{1}{\beta}} \ell_j^{\frac{2}{\beta\wedge 1}} \lesssim V_n^{\frac{1}{\beta}} 3^{jq_1},$$

for a large enough exponent $q_1 \geq 1$.

- **Regime 2: $J_n < j < \infty$.** In this regime, it will suffice to provide a crude bound on $\partial^c \phi_n(L_j)$. We begin with the following result, which is a quantitative generalization of Proposition C.4 of Gangbo and McCann (1996).

  Proposition 9. Let $R, r \geq 4$. Let $\phi : \mathbb{R}^d \to \mathbb{R}$ be a c-concave function such that $|\phi| \leq R$ over $B_{0,r}$. Then, under conditions **(H0)** and **(H3)**, we have

  $$\sup_{y \in \partial^c \phi(B_{0,r/2})} \|y\|^{p-1} \leq C_{p,\kappa}(r^{p-1} + R),$$

  for a constant $C_{p,\kappa} > 0$ depending only on p and κ.

  By Proposition 9, it will suffice to bound $|\phi_n(x)|$ uniformly over $x \in L_j$, for all $j \geq J_n$. Recall from Lemma 24 that $|\phi_n(x)| \leq \overline{R}_n$, thus it suffices to bound $\overline{R}_n$. Define $D_n = \sigma(4 \log n)^{1/\beta}$ and the event $A'_n = \{R_n \leq D_n\}$. Apply a union bound together with the sub-Weibull assumption on P and Q to deduce that

  $$\mathbb{P}((A'_n)^{\mathsf{c}}) \lesssim n \exp\left\{-\frac{1}{2}\left(\frac{D_n}{\sigma}\right)^{\beta}\right\} \lesssim \frac{1}{n}. \tag{3.30}$$

  Over the event A'_n, we therefore have $|\phi_n(x)| \leq \overline{R}_n \lesssim D_n^p$ for all $x \in \mathbb{R}^d$. Combined with Proposition 9, we deduce

  $$\sup_{y \in \partial^c \phi_n(L_j)} \|y\| \lesssim \left[\ell_j + D_n^{\frac{p}{p-1}}\right] \lesssim (\log n)^{\frac{p}{\beta(p-1)}} 3^j \lesssim (3^j \log n)^{q_2},$$

  for a large enough exponent $q_2 > 0$.

In summary, the following holds over the event $A_n \cap A'_n$, uniformly in $j \geq 0$,

$$\sup_{y \in \partial^c \phi_n(L_j)} \|y\| \leq H_j := C_2 \begin{cases} 3^{jq_1} V_n^{\frac{1}{\beta}}, & 0 \leq j \leq J_n \\ (3^j \log n)^{q_2}, & j > J_n. \end{cases}$$

Step 4: Bounding the Lipschitz and Semi-Concavity Moduli of ϕ_n. Define for a large enough constant $C_3 > 0$,

$$\xi_n(x) = C_3 \|x\|^2 \sum_{j=0}^{\infty} H_j^p I(x \in I_j),$$

and set $\widetilde{\phi}_n(x) = \frac{1}{\Lambda}\phi_n(x) - \xi_n(x)$. Under condition **(H3)**, it follows from Lemma 23 that $\widetilde{\phi}_n|_{I_{jk}}$ is $C_4 H_j^p \ell_j$-Lipschitz and concave for all $j \geq 0$ and $k = 1, \dots, m_d$. Furthermore, the map $\widetilde{\phi}_n - \widetilde{\phi}_n(0)$ is bounded over L_j, and hence also over I_{jk}, by $C_5 H_j^p \ell_j^2$. Thus, there exist sufficiently large exponents $r_i \geq 1$, $1 \leq i \leq 4$, such that if $M{=}(M_j)_{j=0}^{\infty}$, $U{=}(U_j)_{j=0}^{\infty}$, where

$$M_j = C \begin{cases} 3^{jr_1}, & 0 \leq j \leq J_n \\ (3^j \log n)^{r_2}, & j > J_n. \end{cases}, \quad U_j = C' \begin{cases} 3^{jr_3}, & 0 \leq j \leq J_n \\ (3^j \log n)^{r_4}, & j > J_n. \end{cases},$$

then $V_n^{-\frac{p}{\beta}}\big(\widetilde{\phi}_n - \widetilde{\phi}_n(0)\big) \in \mathcal{K}_{M,U}$, over the event $A_n \cap A_n'$.

Step 5: Empirical Process Reduction. We deduce from Step 4 that, over $A_n \cap A_n'$,

$$\begin{aligned}\left|\int \phi_n d(P_n - P)\right| &= \Lambda \left|\int \widetilde{\phi}_n d(P_n - P) + \int \xi_n d(P_n - P)\right| \\ &\leq \Lambda \left|\int (\widetilde{\phi}_n - \widetilde{\phi}_n(0)) d(P_n - P)\right| + \Lambda \left|\int \xi_n d(P_n - P)\right| \\ &\leq \Lambda V_n^{\frac{p}{\beta}} \sup_{f \in \mathcal{K}_{M,U}} \int f d(P_n - P) + \Lambda \left|\int \xi_n d(P_n - P)\right|.\end{aligned}$$

Apply the same argument over the event

$$E_n = A_n \cap A_n' \cap \bigcap_{j=0}^{J_n} \left\{ \inf_{x \in I_j} \inf_{\|x-y\| \leq 2} Q_n(B_y) \geq \frac{C_1}{2} \exp(-\ell_j^2/\gamma^2) \right\}$$

to deduce similarly that $V_n^{-p/\beta}(\widetilde{\psi}_n - \widetilde{\psi}_n(0)) \in \mathcal{K}_{M,U}$, where $\widetilde{\psi}_n(y) = \frac{1}{\Lambda}\psi_n(y) - \xi_n(y)$, up to increasing the constants $C, C', C_3 > 0$, and that $\mathbb{P}(E_n^{\mathsf{c}}) \lesssim 1/n$. We thus have, over E_n,

$$\left|\int \psi_n d(Q_n - Q)\right| \leq \Lambda V_n^{\frac{p}{\beta}} \sup_{g \in \mathcal{K}_{M,U}} \int g d(Q_n - Q) + \Lambda \left|\int \xi_n d(Q_n - Q)\right|. \tag{3.31}$$

In the sequel, we write

$$\begin{aligned}\mathcal{X}_n &= \left|\int \xi_n d(P_n - P)\right| + \left|\int \xi_n d(Q_n - Q)\right|, \text{ and,} \\ \Delta_n &= \sup_{f \in \mathcal{K}_{M,U}} \int f d(P_n - P) + \sup_{g \in \mathcal{K}_{M,U}} \int g d(Q_n - Q),\end{aligned}$$

so that,

$$\begin{aligned}\mathbb{E}\left\{ I_{E_n} \left[\left|\int \phi_n d(P_n - P)\right| + \left|\int \psi_n d(Q_n - Q)\right| \right] \right\} &\lesssim \Lambda \mathbb{E}\left[V_n^{\frac{p}{\beta}} \Delta_n \right] + \Lambda \mathbb{E}[\mathcal{X}_n] \\ &\leq \Lambda \left(\mathbb{E}\Big[V_n^{\frac{2p}{\beta}} \Big] \mathbb{E}\left[\Delta_n^2\right] \right)^{1/2} + \Lambda \mathbb{E}[\mathcal{X}_n]\end{aligned}$$

Since P and Q are (σ, β)-sub-Weibull, and since $\rho \geq 2p/\beta$, it readily follows from Jensen's inequality that $\mathbb{E}V_n^{2p/\beta} \leq 2$. Deduce that

$$\mathbb{E}\left\{ I_{E_n} \left[\left|\int \phi_n d(P_n - P)\right| + \left|\int \psi_n d(Q_n - Q)\right| \right] \right\} \lesssim \Lambda \sqrt{\mathbb{E}\left[\Delta_n^2\right]} + \Lambda \mathbb{E}\left[\mathcal{X}_n\right]. \tag{3.32}$$

Step 6: Metric Entropy Bound. Key to bounding $\mathbb{E}[\Delta_n^2]$ is the following upper bound on the $L^2(P_n)$ metric entropy of the class $\mathcal{K}_{M,U}$.

Proposition 10. There exists $C_5 > 0$ such that for all $\epsilon > 0$,

$$\log N(\epsilon, \mathcal{K}_{M,U}, L^2(P_n)) \leq C_5 \cdot V_n^{\frac{d}{4}} \epsilon^{-\frac{2}{d}}.$$

Proof. Using Lemma 20, we prove the following result in Section 3.B.5, inspired by Corollary 2.7.4 of van der Vaart and Wellner (1996).

Lemma 26. *For all* $\epsilon > 0$*,*

$$\log N(\epsilon, \mathcal{K}_{M,U}, L^2(P_n)) \lesssim \left(\frac{1}{\epsilon}\right)^{\frac{d}{2}} \left(\sum_{j=0}^{\infty}\sum_{k=1}^{m_d} (U_j + \mathrm{diam}(I_{jk})M_j)^{\frac{2d}{d+4}} P_n(I_{jk})^{\frac{d}{d+4}}\right)^{\frac{4+d}{4}}.$$

In particular, Lemma 26 implies,

$$\log N(\epsilon, \mathcal{K}_{M,U}, L^2(P_n)) \lesssim \left(\frac{1}{\epsilon}\right)^{\frac{d}{2}} \left(\sum_{j=0}^{\infty} \left(U_j + 3^j M_j\right)^{\frac{2d}{d+4}} P_n(I_j)^{\frac{d}{d+4}}\right)^{\frac{4+d}{4}}.$$

By Markov's inequality, notice that for all $j \geq 0$,

$$P_n(I_j) \leq P_n(B^{\mathsf{c}}_{3^{j-1}}) \lesssim \frac{\int \exp\left(\frac{\|x\|^\beta}{2\rho\sigma^\beta}\right) dP_n(x)}{\exp\left(3^{(j-1)\beta}/2\rho\sigma^\beta\right)} \leq V_n \exp(-c_1 3^{j\beta}),$$

so that

$$\begin{aligned}
\log N(\epsilon, \mathcal{K}_{M,U}, L^2(P_n)) &\lesssim \left(\frac{1}{\epsilon}\right)^{\frac{d}{2}} V_n^{\frac{d}{4}} \left(\sum_{j=0}^{\infty} \left(U_j + 3^j M_j\right)^{\frac{2d}{d+4}} \exp(-c_1 3^{j\beta})\right)^{\frac{4+d}{4}} \\
&\lesssim \left(\frac{1}{\epsilon}\right)^{\frac{d}{2}} V_n^{\frac{d}{4}} \left(\sum_{j=0}^{J_n} \left(3^{jr_1} + 3^j 3^{jr_3}\right)^{\frac{2d}{d+4}} \exp(-c_1 3^{j\beta}) \right.\\
&\qquad \left. + \sum_{j=J_n+1}^{\infty} \left((3^j \log n)^{r_2} + 3^j (3^j \log n)^{r_4}\right)^{\frac{2d}{d+4}} \exp(-c_1 3^{j\beta})\right)^{\frac{4+d}{4}}.
\end{aligned}$$

We deduce that there exist constants $c_2, c_3 > 0$ such that

$$\begin{aligned}
&\log N(\epsilon, \mathcal{K}_{M,U}, L^2(P_n)) \\
&\quad\lesssim \left(\frac{1}{\epsilon}\right)^{\frac{d}{2}} V_n^{\frac{d}{4}} \left[\sum_{j=0}^{J_n} 3^{jc_2} \exp(-c_3 3^{j\beta}) + (\log n)^{c_2} \sum_{j=J_n+1}^{\infty} 3^{jc_2} \exp(-c_3 3^{j\beta})\right]^{\frac{4+d}{4}}.
\end{aligned}$$

Notice that $\sum_{j=0}^{\infty} 3^{jc_2} \exp(-c_3 3^{j\beta}) < \infty$, thus the first summation on the right-hand side of the above display is finite. For the second summation, notice that there exists $J_0 > 0$ such that for all $j \geq J_0$, $3^{jc_2} \leq \exp(c_4 3^{j\beta})$ where $c_4 = c_3/2$. Thus, since $J_n \asymp \log\log n$, we obtain

$$\log N(\epsilon, \mathcal{K}_{M,U}, L^2(P_n))$$

$$\lesssim \left(\frac{1}{\epsilon}\right)^{\frac{d}{2}} V_n^{\frac{d}{4}} \left[1 + (\log n)^{c_2} \sum_{j=J_n+1}^{\infty} \exp(-c_4 3^{j\beta})\right]^{\frac{4+d}{4}}$$

$$\lesssim \left(\frac{1}{\epsilon}\right)^{\frac{d}{2}} V_n^{\frac{d}{4}} \Big[1 + (\log n)^{c_2} \exp\Big(- c_4(3^{(J_n+1)\beta} + 1)\Big)\Big]^{\frac{4+d}{4}} \lesssim \left(\frac{1}{\epsilon}\right)^{\frac{d}{2}} V_n^{\frac{d}{4}}.$$

This proves Proposition 10. □

Step 7: Chaining. Equipped with Proposition 10, we are now in a position to bound the expected (squared) supremum of the empirical process indexed by $\mathcal{K}_{M,U}$. We begin by noting the following.

Lemma 27. *It holds that*

$$\mathbb{E}\left[\left(\sup_{f\in\mathcal{K}_{M,U}} \int f d(P_n - P)\right)^2\right] \lesssim \frac{(\log n)^{2r_4}}{n} + \mathbb{E}\left[\sup_{f\in\mathcal{K}_{M,U}} \int f d(P_n - P)\right]^2.$$

By combining Lemma 27 with Lemma 19 and Proposition 10, we deduce that for all $\tau > 0$,

$$\mathbb{E}\left[\left(\sup_{f\in\mathcal{K}_{M,U}} \int f d(P_n - P)\right)^2\right] \lesssim \frac{(\log n)^{2r_4}}{n} + \left(\tau + \frac{\mathbb{E}V_n^{\frac{d}{4}}}{\sqrt{n}} \int_\tau^\infty \left(\frac{1}{\epsilon}\right)^{\frac{d}{4}} d\epsilon\right)^2.$$

Since P and Q are (σ,β)-sub-Weibull, and since $\rho \geq d/4$, we again have by Jensen's inequality that $\mathbb{E}[V_{i,n}^{d/4}] \leq 2$ for both $i = 1,2$, implying that $\mathbb{E}[V_n^{d/4}] \leq c_5$. Choosing $\tau \asymp n^{-2/d}$ in the above display thus leads to a bound scaling at the rate $n^{-4/d}$. Upon repeating the same argument for Q_n, we obtain

$$\sqrt{\mathbb{E}[\Delta_n^2]} \lesssim n^{-2/d}. \tag{3.33}$$

Step 8: Conclusion. Let $(\phi_0,\psi_0) \in \Phi_c(P,Q)$ be a pair of optimal Kantorovich potentials between P and Q. It follows similarly as in the proof of Theorem 9 that

$$\mathcal{T}_c(P_n,Q_n) - \mathcal{T}_c(P,Q) \geq \int \phi_0 d(P_n - P) + \int \psi_0 d(Q_n - Q) =: \Gamma_n.$$

Combine the above display with equations (3.27), (3.32) and (3.33) to deduce

$$\begin{aligned}
&\mathbb{E}\big|\mathcal{T}_c(P_n,Q_n) - \mathcal{T}_c(P,Q)\big| \\
&\quad\leq \mathbb{E}|\Gamma_n| + \mathbb{E}\left[\int \phi_n d(P_n - P)\right] + \mathbb{E}\left[\int \psi_n d(Q_n - Q)\right] \\
&\quad\lesssim \mathbb{E}|\Gamma_n| + \Lambda\Big\{\mathbb{E}[\mathcal{X}_n] + \sqrt{\mathbb{E}[\Delta_n^2]}\Big\} + \mathbb{E}\left[I_{E_n^c}\int \phi_n d(P_n - P)\right] + \mathbb{E}\left[I_{E_n^c}\int \psi_n d(Q_n - Q)\right] \\
&\quad\lesssim \mathbb{E}|\Gamma_n| + \Lambda\Big\{\mathbb{E}[\mathcal{X}_n] + \sqrt{\mathbb{E}[\Delta_n^2]}\Big\} + \sqrt{\mathbb{P}(E_n^c)\mathbb{E}[\overline{R}_n^2]} \\
&\quad\lesssim \mathbb{E}|\Gamma_n| + \Lambda\Big\{\mathbb{E}[\mathcal{X}_n] + n^{-2/d}\Big\} + n^{-1/2}(\log n)^{c_1}.
\end{aligned} \tag{3.34}$$

The quantities Γ_n and $\mathcal{X}_n$ are simple to bound in expectation, as we now show.

Lemma 28. *For any* $\epsilon > 0$, $\mathbb{E}|\Gamma_n| \vee \mathbb{E}[\mathcal{X}_n] \lesssim n^{\epsilon-\frac{1}{2}}$.

Since $d \geq 5$, combining equation (3.34) with Lemma 28 leads to the claim. □

3.3.2 Proof of Theorem 10

Fix $x \in \mathbb{R}^d$. If $\partial^c\phi(x)$ is empty, then there is nothing to show, thus suppose otherwise. Choose $y_x \in \partial^c\phi(x)$. Let $K_p = 2(3\kappa^2)^{1/(p-1)}$, and notice that we may assume

$$\|y_x\| \geq 4(K_p+1)(\|x\|+1), \tag{3.35}$$

as otherwise we are done. In particular, this assumption implies $\|y_x - x\| \geq 4$, thus the following ball is non-empty

$$U = \left\{ u \in \mathbb{R}^d : \|u - y_x\| \leq \|x - y_x\| - 1 \right\}.$$

Furthermore, define $\xi = (y_x - x)/\|y_x - x\|$, and note that the ball $S := B_{x+2\xi}$ of radius 1 centered at $x + 2\xi$ is contained in U.

If $\partial^c\phi(S) = \emptyset$, then the condition $\mathrm{supp}(\pi) \subseteq \partial^c\phi$ implies $P(S) = 0$, in which case the right-hand side of equation (3.24) is infinite and the claim is trivial. Thus, assume otherwise, and pick $u \in S$ for which $\partial^c\phi(u)$ is nonempty. Notice that $\|x - u\| \leq 3$. Furthermore, let $y_u \in \partial^c\phi(u)$ be arbitrary. Since ϕ is c-concave, the set $\partial^c\phi$ is c-cyclically monotone by Lemma 17(iii). In particular,

$$c(x, y_x) - c(u, y_x) \leq c(x, y_u) - c(u, y_u).$$

Thus, using condition (**H3**),

$$\begin{aligned}
c(x,y_u) - c(u,y_u) &\geq \omega(\|x-y_x\|) - \omega(\|u-y_x\|) && (\text{Since } \|x-y_x\| \wedge \|u-y_x\| \geq 1)\\
&\geq \omega(\|x-y_x\|) - \omega(\|x-y_x\|-1) && (\text{Since } u \in U \text{ and } \omega \text{ is increasing})\\
&\geq \omega'(\|x-y_x\|-1) && (\text{By convexity of } \omega)\\
&\geq \frac{1}{\kappa}(\|x-y_x\|-1)^{p-1} && (\text{By condition } (\mathbf{H3}))\\
&\geq \frac{1}{\kappa 2^{p-1}}\|x-y_x\|^{p-1}. && (\text{Since } \|x-y_x\| \geq 4)
\end{aligned}$$

Now, conditions (**H0**) and (**H3**) imply that $h(z) \leq \kappa$ for all $z \in B_{0,1}$. Furthermore, the preceding display combined with equation (3.35) implies that $c(x, y_u) > \kappa$, whence $\|x - y_u\| \geq 1$. We may thus again apply conditions (**H0**) and (**H3**) to obtain,

$$c(x,y_u) - c(u,y_u) \leq \langle \nabla h(x-y_u), x-u\rangle \leq \omega'(\|x-y_u\|)\|x-u\| \leq 3\kappa\|x-y_u\|^{p-1}.$$

We deduce that $\|x - y_x\| \leq K_p\|x - y_u\|$, whence,

$$\|y_x\| \leq K_p\|x-y_u\| + \|x\| \leq K_p\|y_u\| + \|x\|(K_p+1) \leq K_p\|y_u\| + \frac{1}{4}\|y_x\|,$$

where the last inequality is due to equation (3.35). We thus have $\|y_u\| \geq C\|y_x\|$ for a constant $C > 0$ depending only on d, p, κ. It follows that,

$$\partial^c\phi(S) \subseteq \left\{v \in \mathbb{R}^d : \|v\| \geq C\|y_x\|\right\}.$$

Given $Y \sim Q$, we deduce from the sub-Weibull condition on Q that

$$Q(\partial^c\phi(S)) \leq \mathbb{P}\left(\|Y\| \geq C\|y_x\|\right) \lesssim \exp\left(-\frac{C^\beta \|y_x\|^\beta}{2\sigma^\beta}\right).$$

On the other hand, using the fact that $\mathrm{supp}(\pi) \subseteq \partial^c\phi$, one has

$$Q(\partial^c\phi(S)) = \pi(\mathbb{R}^d \times \partial^c\phi(S)) \geq \pi(S \times \partial^c\phi(S)) = P(S),$$

so that,

$$\exp\left(-\frac{C^\beta \|y_x\|^\beta}{2\sigma^\beta}\right) \gtrsim Q(\partial^c\phi(S)) \geq P(S) \geq \inf_{y:\|x-y\|\leq 2} P(B_y). \tag{3.36}$$

The first claim follows. To prove the second claim, recall from the definition of (γ, b)-super-Gaussianity that for all $y \in \mathbb{R}^d$ such that $\|x - y\| \leq 2$,

$$P(B_y) \geq \frac{b}{\sqrt{2\pi\gamma^2}}\mathcal{L}(B_y) \inf_{z\in B_y} \exp(-\|z\|^2/2\gamma^2) \geq C_1 \exp(-\|x\|^2/C_1),$$

for a constant $C_1 > 0$ depending on d, b, γ. By Lemma 17(iv), since $\mathcal{T}_c(P,Q) < \infty$, any optimal coupling π between P and Q lies in the support of a c-concave potential ϕ. We may therefore apply equation (3.36) to deduce that, for some constant $C' > 0$, any such coupling satisfies

$$\|y\| \leq C'\sigma(\|x\| + 1)^{\frac{2}{\beta}}, \quad \pi\text{-a.e. } (x, y).$$

Furthermore, since P is absolutely continuous with respect to the Lebesgue measure, notice that the conditions of Gangbo and McCann (1996), Theorem 1.2, are satisfied under conditions (**H0**) and (**H3**) and the strict convexity of h. Therefore, there exists a unique optimal transport map T from P to Q, so that the measure π in the above display may be taken to be $(Id, T)_{\#}P$. The claim follows. □

3.4 Lower Bounds

In this section, we derive two lower bounds which imply that the rates of convergence derived in Sections 3.2 and 3.3 are typically unimprovable. In Section 3.4.1, we obtain lower bounds on the rate of convergence of the empirical optimal transport cost, while in Section 3.4.2, we derive a minimax lower bound which implies that, up to polylogarithmic factors, no estimator of $\mathcal{T}_c(P,Q)$ can achieve a faster rate of convergence than the empirical estimator uniformly over all pairs of measures P, Q.

In order to state our lower bounds, we require an assumption on the maximal Hölder exponent $\alpha \in (0, 2]$ achievable by the cost h. To state such an assumption, recall that our upper bounds were based, for instance, on the condition $\Lambda = 1 \vee \|h\|_{\mathscr{C}^\alpha(\mathcal{Z}_1)} < \infty$, for some $\alpha \in (0, 2]$, which in particular implies that for all $z, z_0 \in \mathcal{Z}$,

$$h(z) - h(z_0) \leq \begin{cases} \Lambda \|z - z_0\|^\alpha, & \alpha \leq 1 \\ \langle \nabla h(z_0), z - z_0 \rangle + \Lambda \|z - z_0\|^\alpha, & \alpha > 1 \end{cases}.$$

We shall assume the following dual condition throughout this section.

(**H4**) $\mathcal{X}$ and $\mathcal{Y}$ are convex sets with nonempty interior, and are such that h is differentiable over $\mathcal{Z} = \mathcal{X} - \mathcal{Y}$. Furthermore, there exist $\lambda > 0$, $\alpha \in (0, 2]$, and $z_0 = x_0 - y_0 \in \mathcal{Z}$ such that $x_0 \in \mathrm{int}(\mathcal{X})$, $y_0 \in \mathrm{int}(\mathcal{Y})$, and for all $z \in \mathcal{Z}$,

$$h(z) - h(z_0) \geq \begin{cases} \lambda \|z - z_0\|^\alpha, & \alpha \leq 1 \\ \langle \nabla h(z_0), z - z_0 \rangle + \lambda \|z - z_0\|^\alpha, & \alpha > 1 \end{cases}.$$

Notice that condition (**H4**) implies that $\mathcal{X}$ and $\mathcal{Y}$ have positive Lebesgue measure over $\mathbb{R}^d$. The presence of this assumption can be anticipated from the fact that empirical optimal transport costs may achieve improved rates of convergence when $\mathcal{X}$ and $\mathcal{Y}$ have intrinsic dimension less than d (Weed and Bach, 2019). It is straightforward to verify that conditions (**H0**) and (**H4**) are satisfied by the cost $h(x) = \|x\|^p$ when $1 \leq p \leq 2$ with $\alpha = p$ and $z_0 = 0$. These conditions are also satisfied for $2 < p < \infty$ and $\alpha = 2$, whenever there exists a neighborhood of zero which is not contained in $\mathcal{Z}$. We also note that condition (**H4**) is satisfied with $\alpha = 2$ by any differentiable and (2λ)-strongly convex function h over $\mathcal{Z}$.

Finally, we assume throughout this section that X_i is independent of Y_j for all $1 \leq i, j \leq n$. Though this condition is not needed to derive our upper bounds, we do not preclude the possibility that they may be sharpened under particular dependence structures between the samples from P and Q.

3.4.1 Lower Bounds for the Empirical Optimal Transport Cost

We begin with the following lower bound on the rate of convergence of the empirical optimal transport cost.

Proposition 11. Assume conditions (**H0**) and (**H4**). Then,

$$\sup_{\substack{P \in \mathcal{P}(\mathcal{X}) \\ Q \in \mathcal{P}(\mathcal{Y})}} \mathbb{E}_{P,Q} \big| \mathcal{T}_c(P_n, Q_n) - \mathcal{T}_c(P, Q) \big| \gtrsim \lambda n^{-\alpha/d}.$$

Proposition 11 implies that the upper bounds in Theorems 9, 11, and Corollaries thereafter cannot generally be improved, provided that the Hölder exponent α therein is chosen maximally in the sense of condition (**H4**). As we now show, our lower bound is constructive, and is typically achieved by absolutely continuous measures differing by a location translation, as in Example 3.1.1.

Proof. We prove the claim assuming $\alpha \in (1,2]$, and an analogous argument may be used when $\alpha \in (0,1]$. Since $x_0 \in \mathrm{int}(\mathcal{X})$ and $y_0 \in \mathrm{int}(\mathcal{Y})$, there exists $\epsilon > 0$ such that $\mathcal{X}_0 := B_{x_0,\epsilon} \subseteq \mathcal{X}$ and $\mathcal{Y}_0 := B_{y_0,\epsilon} \subseteq \mathcal{Y}$. Define the measures

$$P = \frac{\mathcal{L}|_{\mathcal{X}_0}}{\mathcal{L}(\mathcal{X}_0)}, \quad Q = \frac{\mathcal{L}|_{\mathcal{Y}_0}}{\mathcal{L}(\mathcal{Y}_0)},$$

where recall that $\mathcal{L}$ is the Lebesgue measure on $\mathbb{R}^d$. By construction, $Q = T_{0\#}P$ where $T_0(x) = x + z_0$. Since h is convex, it follows by the same argument as in Example 3.1.1 that T_0 is an optimal transport map from P to Q.

Let γ_n denote an optimal coupling between P_n and Q with respect to the cost c. Then, by condition **(H4)**,

$$\begin{aligned}
\mathcal{T}_c(P_n, Q) - \mathcal{T}_c(P, Q) &= \int \Big[c(x,y) - c(x, T_0(x))\Big] d\gamma_n(x,y) \\
&= \int \Big[h(y-x) - h(z_0)\Big] d\gamma_n(x,y) \\
&\geq \int \Big[\langle \nabla h(z_0), y - x - z_0\rangle + \lambda \left\|y - x - z_0\right\|^\alpha \Big] d\gamma_n(x,y) \\
&= \int \Big[\langle \nabla h(z_0), y - x\rangle + \lambda \left\|y - x\right\|^\alpha \Big] d\pi_n(x,y),
\end{aligned}$$

where $\pi_n = (Id, T_0^{-1})_\# \gamma_n \in \Pi(P_n, P)$. It follows that

$$\begin{aligned}
\mathcal{T}_c(P_n, Q) - \mathcal{T}_c(P, Q) &\geq \int \langle \nabla h(z_0), y - x\rangle d\pi_n(x,y) + \lambda W_\alpha^\alpha(P_n, P) \\
&\geq \int \langle \nabla h(z_0), \cdot\rangle d(P - P_n) + \lambda W_1^\alpha(P_n, P).
\end{aligned}$$

The first order term on the final line of the above display clearly has mean zero, whence

$$\mathbb{E}\Big[\mathcal{T}_c(P_n, Q) - \mathcal{T}_c(P, Q)\Big] \geq \lambda \mathbb{E} W_1^\alpha(P_n, P) \geq \lambda \big[\mathbb{E} W_1(P_n, P)\big]^\alpha \gtrsim \lambda n^{-\alpha/d},$$

where the final inequality follows from Proposition 2.1 of Dudley (1969), due to the absolute continuity of P with respect to the Lebesgue measure on $\mathbb{R}^d$. Finally, since h is continuous and $\mathcal{Z}_0 = \mathcal{X}_0 - \mathcal{Y}_0$ is compact, h is bounded over $\mathcal{Z}_0$. Thus, by Lemma 17(ii), there exists a pair of Kantorovich potentials (ϕ_n, ψ_n) such that $\mathcal{T}_c(P_n, Q) = J_{P_n, Q}(\phi_n, \psi_n)$, whence

$$\mathcal{T}_c(P_n, Q_n) \geq J_{P_n, Q_n}(\phi_n, \psi_n) = \mathcal{T}_c(P_n, Q) + \int \psi_n d(Q_n - Q).$$

Since the random variables $X_1, \dots, X_n$ are independent of $Y_1, \dots, Y_n$, ψ_n is also independent of $Y_1, \dots, Y_n$, whence

$$\mathbb{E}\left[\int \psi_n d(Q_n - Q)\middle| X_1, \dots, X_n\right] = 0.$$

It readily follows that $\mathbb{E}\mathcal{T}_c(P_n, Q_n) \geq \mathbb{E}\mathcal{T}_c(P_n, Q)$, so that

$$\mathbb{E}\big|\mathcal{T}_c(P_n, Q_n) - \mathcal{T}_c(P, Q)\big| \geq \mathbb{E}\big[\mathcal{T}_c(P_n, Q_n) - \mathcal{T}_c(P, Q)\big] \geq \mathbb{E}\big[\mathcal{T}_c(P_n, Q) - \mathcal{T}_c(P, Q)\big] \gtrsim \lambda n^{-\alpha/d}.$$

The claim follows. □

3.4.2 Minimax Lower Bounds

We next turn to deriving a minimax lower bound on the rate of estimating the optimal transport cost between two probability measures. Unlike Proposition 11, our next result will require both condition (**H4**) and the smoothness condition (**H1**) from Section 3.2.

Theorem 12. *Assume conditions (**H0**), (**H1**) and (**H4**). Then, there exists a constant $C > 0$ depending on $\lambda, \Lambda, d, \mathcal{X}, \mathcal{Y}, \alpha$ such that*

$$\inf_{\widehat{\mathcal{T}}_n} \sup_{\substack{P \in \mathcal{P}(\mathcal{X}) \\ Q \in \mathcal{P}(\mathcal{Y})}} \mathbb{E}_{P,Q} \left| \widehat{\mathcal{T}}_n - \mathcal{T}_c(P,Q) \right| \geq C(n \log n)^{-\alpha/d},$$

where the infimum is over all Borel-measurable functions $\widehat{\mathcal{T}}_n$ of $X_1, \dots, X_n$ and $Y_1, \dots, Y_n$.

Theorem 12 shows that the convergence rates exhibited throughout this chapter for the empirical optimal transport cost estimator cannot be improved by any other estimator uniformly over all pairs of measures in $\mathcal{P}(\mathcal{X}) \times \mathcal{P}(\mathcal{Y})$, up to a polylogarithmic factor. Minimax lower bounds scaling at the rate $(n \log n)^{-1/d}$ have previously been established for the problem of estimating p-Wasserstein distances by Niles-Weed and Rigollet (2022) (Theorem 11), and we build upon their techniques to prove Theorem 12. The proof is deferred to Section 3.C.

3.A Omitted Proofs from Section 3.2

3.A.1 Proof of Corollary 3

Throughout the proof, $C > 0$ denotes a constant depending only on d, p, r, whose value may change from line to line.

The proof is elementary, but tedious. To prove the first claim, it suffices to show that $\|\cdot\|_{\ell_r}^p \in \mathscr{C}^{2 \wedge r \wedge p}(B_{0,2})$. It is clear that $\|\cdot\|_{\ell_r}$ and $\|\cdot\|_{\ell_1}^p$ are Lipschitz for any $r, p \geq 1$, thus it suffices to assume $p, r > 1$. In this case, $\|\cdot\|_{\ell_r}^p$ is differentiable, and for all $l = 1, \dots, d$,

$$\frac{\partial \|x\|_{\ell_r}^p}{\partial x_l} = p x_l |x_l|^{r-2} \|x\|_{\ell_r}^{p-r}. \tag{3.37}$$

Next, we show that $\frac{\partial \|\cdot\|_{\ell_r}^p}{\partial x_l}$ is Hölder continuous over $B_{0,2}$ with suitable exponent, uniformly in l. Let $x, y \in B_{0,2}$, and assume without loss of generality that $\|x\|_{\ell_r} \leq \|y\|_{\ell_r}$. Then,

$$\begin{aligned} &\left| x_l |x_l|^{r-2} \|x\|_{\ell_r}^{p-r} - y_l |y_l|^{r-2} \|y\|_{\ell_r}^{p-r} \right| \\ &\qquad \leq \left| x_l |x_l|^{r-2} \|x\|_{\ell_r}^{p-r} - x_l |x_l|^{r-2} \|y\|_{\ell_r}^{p-r} \right| + \left| x_l |x_l|^{r-2} \|y\|_{\ell_r}^{p-r} - y_l |y_l|^{r-2} \|y\|_{\ell_r}^{p-r} \right| \\ &\qquad = |x_l|^{r-1} \left| \|x\|_{\ell_r}^{p-r} - \|y\|_{\ell_r}^{p-r} \right| + \|y\|_{\ell_r}^{p-r} \left| x_l |x_l|^{r-2} - y_l |y_l|^{r-2} \right|. \end{aligned}$$

For the first term, notice that

$$|x_l|^{r-1} \left| \|x\|_{\ell_r}^{p-r} - \|y\|_{\ell_r}^{p-r} \right| \leq \|x\|_{\ell_r}^{r-1} \left(\|y\|_{\ell_r}^{p-r} - \|x\|_{\ell_r}^{p-r} \right)$$

$$\le \|y\|_{\ell_r}^{p-1} - \|x\|_{\ell_r}^{p-1} \le C\,\|y-x\|^{1\wedge(p-1)}.$$

Furthermore, letting $\epsilon_x = \mathrm{sgn}(x_l)$ and $\epsilon_y = \mathrm{sgn}(y_l)$, we have,

$$\begin{aligned}
&\|y\|_{\ell_r}^{p-r}\left|x_l|x_l|^{r-2} - y_l|y_l|^{r-2}\right| \\
&\qquad= \|y\|_{\ell_r}^{p-r}\left|\epsilon_x|x_l|^{r-1} - \epsilon_y|y_l|^{r-1}\right| \\
&\qquad\le \|y\|_{\ell_r}^{p-r}\left(\left|\epsilon_x|x_l|^{r-1} - \epsilon_x|y_l|^{r-1}\right| + \left|\epsilon_x|y_l|^{r-1} - \epsilon_y|y_l|^{r-1}\right|\right) \\
&\qquad\le \|y\|_{\ell_r}^{p-r}\left(\left||x_l|^{r-1} - |y_l|^{r-1}\right| + |y_l|^{r-1}\left|\epsilon_x - \epsilon_y\right|\right) \\
&\qquad\le \|y\|_{\ell_r}^{p-r}\left(\left||x_l|^{r-1} - |y_l|^{r-1}\right| + 2|x_l - y_l|^{r-1}\right).
\end{aligned} \tag{3.38}$$

We now consider several cases. If $p \ge r$, then we readily obtain

$$\|y\|_{\ell_r}^{p-r}\left|x_l|x_l|^{r-2} - y_l|y_l|^{r-2}\right| \le C\left(|x_l - y_l|^{1\wedge(r-1)} + 2|x_l-y_l|^{r-1}\right) \le C|x_l - y_l|^{1\wedge(r-1)}.$$

If instead $p < r \le 2$, then from equation (3.38), we obtain

$$\begin{aligned}
\|y\|_{\ell_r}^{p-r}\left|x_l|x_l|^{r-2} - y_l|y_l|^{r-2}\right| &\le 3\,\|y\|_{\ell_r}^{p-r}|x_l - y_l|^{r-1} \\
&\le 3\,\|y\|_{\ell_r}^{p-r}(|x_l| + |y_l|)^{r-p}|x_l - y_l|^{p-1} \le C|x_l - y_l|^{p-1}.
\end{aligned}$$

Finally, if $p < r$ and $r > 2$, we have from equation (3.38),

$$\begin{aligned}
\|y\|_{\ell_r}^{p-r}\left|x_l|x_l|^{r-2} - y_l|y_l|^{r-2}\right| &\le \|y\|_{\ell_r}^{p-r}\left((r-1)(|x_l|\vee|y_l|)^{r-2}|x_l - y_l| + 2|x_l - y_l|^{r-1}\right) \\
&\le C\,\|y\|_{\ell_r}^{p-r}(|x_l| + |y_l|)^{r-2}|x_l - y_l| \\
&\le C\,\|y\|_{\ell_r}^{p-r}(|x_l| + |y_l|)^{r-p}|x_l - y_l|^{p-1} \le C|x_l - y_l|^{p-1}.
\end{aligned}$$

The preceding displays readily imply that for all l, $\partial\,\|\cdot\|_{\ell_r}^p/\partial x_l \in \mathscr{C}^{1\wedge(r-1)\wedge(p-1)}(B_{0,2})$, implying that $\|\cdot\|_{\ell_r}^p \in \mathscr{C}^{2\wedge r\wedge p}(B_{0,2})$. The first claim now follows by applying Theorem 9. To prove the second claim, notice that when $r \ge 2$, the map in equation (3.37) is differentiable with respect to x_l over any open subset of $B_{0,2}$ which does not contain the origin, with derivative

$$\frac{\partial^2\,\|x\|_{\ell_r}^p}{\partial x_l^2} = p\frac{\partial}{\partial x_l}x_l|x_l|^{r-2}\,\|x\|_{\ell_r}^{p-r} = p\,\|x\|_{\ell_r}^{p-2r}\,|x_l|^{r-2}\left[(p-r)|x_l|^r + (r-1)\,\|x\|_{\ell_r}^r\right].$$

Recall that for all positive semidefinite matrices $A \in \mathbb{R}^{d\times d}$, the 1-Schatten norm of A is equal to its trace, so that $\|A\|_\infty \lesssim \mathrm{tr}(A)$. Thus, for any $\epsilon > 0$,

$$\sup_{x\in B_{0,2}\setminus B_{0,\epsilon}} \left\|\nabla^2\,\|x\|_{\ell_r}^p\,\right\|_\infty \lesssim \sup_{x\in B_{0,2}\setminus B_{0,\epsilon}} \sum_{l=1}^d \left|\frac{\partial^2\,\|x\|_{\ell_r}^p}{\partial x_l^2}\right| \lesssim \sup_{x\in B_{0,2}\setminus B_{0,\epsilon}} \sum_{l=1}^d \|x\|_{\ell_r}^{p-r}\,|x_l|^{r-2} < \infty.$$

It readily follows that $\|\cdot\|_{\ell_r}^p \in \mathscr{C}^2(B_{0,2}\setminus B_{0,\epsilon}^\circ)$, with Hölder norm depending only on d, r, p, ϵ. Now, since $\mathcal{X}, \mathcal{Y}$ are convex by condition (**S1**), so is the set $\mathcal{Z} = \mathcal{X} - \mathcal{Y}$, and since $\mathcal{X}$ and $\mathcal{Y}$ are closed and disjoint, there must exist $\epsilon > 0$ such that $B_{0,\epsilon} \cap \mathcal{Z} = \emptyset$. Choose any convex open set $\mathcal{Z}_1$ containing $\mathcal{Z}$, and contained in $B_{0,2}\setminus B_{0,\epsilon/2}^\circ$, to deduce that condition (**H1**) holds with $\alpha = 2$, and with Λ depending only on $d, p, r, \mathcal{X}, \mathcal{Y}$. The claim follows. □

3.A.2 Proof of Lemma 18

The proof is analogous to that of Proposition C.2 of (Gangbo and McCann, 1996), and is included for completeness. We prove the claim for $\widetilde{\phi}_n$, noting that a symmetric argument can be used for the map $\widetilde{\psi}_n$. Define the modified cost function

$$h_\Lambda : z \in \mathcal{Z} \mapsto h(z) - \frac{\Lambda}{2}\|z\|^2 .$$

By condition (**H1**) with $\alpha = 2$, ∇h is Λ-Lipschitz over $\mathcal{Z}$, implying that for all $z_1, z_2 \in \mathcal{Z}$,

$$\langle \nabla h_\Lambda(z_1) - \nabla h_\Lambda(z_2), z_1 - z_2\rangle = \langle \nabla h(z_1) - \nabla h(z_2), z_1 - z_2\rangle - \Lambda\|z_1 - z_2\|^2 \leq 0.$$

It follows that $-\nabla h_\Lambda$ is monotone, whence h_Λ is concave (Hiriart-Urruty and Lemaréchal (2004), Theorem 4.1.4). Now, notice that for all $x \in \mathcal{X}$,

$$\begin{aligned}
\widetilde{\phi}_n(x) &= \inf_{y\in\mathcal{Y}} \{c(x,y) - \psi_n(y)\} - \frac{\Lambda}{2}\|x\|^2 \\
&= \inf_{y\in\mathcal{Y}} \left\{ h(x-y) - \frac{\Lambda}{2}\Big[\|x-y\|^2 - \|y\|^2 + 2\langle x,y\rangle\Big] - \psi_n(y)\right\} \\
&= \inf_{y\in\mathcal{Y}} \left\{ h_\Lambda(x-y) + \frac{\Lambda}{2}\Big[\|y\|^2 - 2\langle x,y\rangle\Big] - \psi_n(y)\right\}.
\end{aligned}$$

By concavity of h_Λ, the last line of the above display is an infimum of concave functions of x. It follows that $\widetilde{\phi}_n$ is concave. To prove that $\widetilde{\phi}_n$ is Lipschitz, let $x \in \mathcal{X}$ and let $(y_k) \subseteq \mathcal{Y}$ be a sequence such that

$$\widetilde{\phi}_n(x) \geq c(x,y_k) - \psi_n(y_k) - \frac{\Lambda}{2}\|x\|^2 - k^{-1}.$$

Then, for all $x' \in \mathcal{X}$ and $k \geq 1$,

$$\begin{aligned}
\widetilde{\phi}_n(x') - \widetilde{\phi}_n(x) &\leq \left[c(x',y_k) - \psi_n(y_k) - \frac{\Lambda}{2}\|x'\|^2\right] - \left[c(x,y_k) - \psi_n(y_k) - \frac{\Lambda}{2}\|x\|^2\right] + k^{-1} \\
&= h(x'-y_k) - h(x-y_k) - \frac{\Lambda}{2}\left[\|x'\|^2 - \|x\|^2\right] + k^{-1} \\
&\leq \left(\sup_{z\in\mathcal{Z}} \|\nabla h(z)\|\right)\|x'-x\| - \frac{\Lambda}{2}(\|x'\| - \|x\|)(\|x'\| + \|x\|) + k^{-1} \\
&\leq \left(\sup_{z\in\mathcal{Z}} \|\nabla h(z)\| + \Lambda\right)\|x'-x\| + k^{-1} \leq 2\Lambda\|x'-x\| + k^{-1}.
\end{aligned}$$

Since k is arbitrary, the Lipschitz property follows upon repeating a symmetric argument to upper bound $\widetilde{\phi}_n(x) - \widetilde{\phi}_n(x')$. Finally, since $|\phi_n| \leq 1$, $|\widetilde{\phi}_n| \leq 2\Lambda$ as $\Lambda \geq 1$. □

3.A.3 Proof of Lemma 21

To prove the first part, recall that condition (**H1**) implies $h \in \mathscr{C}^\alpha(\mathcal{Z}_1)$ with $1 < \alpha < 2$, and $\Lambda \geq \|h\|_{\mathscr{C}^\alpha(\mathcal{Z}_1)}$. For any $z \in \mathcal{Z}$, let $A_z := \{u \in \mathbb{R}^d : z - u \in \mathcal{Z}_1\}$. Since $\mathcal{Z}$ is compact and

$\mathcal{Z}_1$ is open, there exists $\epsilon > 0$ such that for all $z \in \mathcal{Z}$, $B_{0,\epsilon} \subseteq A_z$. In particular, if $\sigma < \epsilon$, then $u \in A_z$ for all $z \in \mathcal{Z}$ and u in the support of K_σ.

Moreover, we have by a first-order Taylor expansion that for all $z \in \mathcal{Z}$ and $u \in A_z$,

$$h(z-u) - h(z) = -\langle \nabla h(z-tu), u \rangle,$$

for some $t \in (0,1)$. By convexity of $\mathcal{Z}_1$, we have $z - tu \in \mathcal{Z}_1$. It follows that

$$|h(z-u) - h(z) + \langle \nabla h(z), u\rangle| \le |\langle \nabla h(z) - \nabla h(z-tu), u\rangle| \le \Lambda \|u\|^\alpha .$$

Finally, the fact that K is even implies that $\int uK_\sigma(u)du = 0$. Combining these facts, we obtain

$$\begin{aligned} |h_\sigma(z) - h(z)| &= \left| \int \big[h(z-u) - h(z)\big] K_\sigma(u)du \right| \\ &\le \left| \int \big[h(z-u) - h(z) + \langle \nabla h(z), u\rangle\big] K_\sigma(u)du \right| + \left| \int \langle \nabla h(z), u\rangle K_\sigma(u)du \right| \qquad (3.39) \\ &\le \Lambda \int \|u\|^\alpha K_\sigma(u)du \le \Lambda\sigma^\alpha , \qquad (3.40) \end{aligned}$$

since the support of K_σ lies in $B_{0,\sigma}$. This proves the first claim.

To prove the second part, it is easy to see that the cost h_σ is convex, even, and lower semi-continuous by assumption on h, thus h_σ satisfies assumption (**H0**). Now, let $\widetilde{\mathcal{Z}}_1$ be an open set such that $\mathcal{Z} \subseteq \widetilde{\mathcal{Z}}_1$ and such that $\mathrm{cl}(\widetilde{\mathcal{Z}}_1) \subseteq \mathcal{Z}_1$. After possibly decreasing the value of $\epsilon > 0$, we may again ensure that $B_{z,\epsilon} \subseteq \mathcal{Z}_1$ for all $z \in \widetilde{\mathcal{Z}}_1$. We shall now prove that h_σ satisfies assumption (**H1**) with the Hölder norm $\|h_\sigma\|_{\mathscr{C}^2(\widetilde{\mathcal{Z}}_1)} \le C\Lambda\sigma^{\alpha-2}$ as long as $\sigma < \epsilon$. That $h_\sigma \le 1$ on $\widetilde{\mathcal{Z}}_1$ is immediate, so it suffices to show that h_σ has the requisite Hölder norm.

Define for any given $z \in \mathcal{Z}$ and all $u \in \mathbb{R}^d$,

$$\widetilde{h}(u) = h(u) - h(z) - \langle \nabla h(z), u - z\rangle, \quad \widetilde{h}_\sigma = \widetilde{h} \star K_\sigma.$$

As before, for any $z \in \widetilde{\mathcal{Z}}_1$ and any $u \in \mathbb{R}^d$ such that $\|u - z\| \le \epsilon$, we have $u \in \mathcal{Z}_1$, whence a first-order Taylor expansion leads to

$$|\widetilde{h}(u)| \le \Lambda \|z - u\|^\alpha .$$

We thus obtain for all $z \in \widetilde{\mathcal{Z}}_1$,

$$\begin{aligned} \|\nabla^2 h_\sigma(z)\|_\infty &= \|\nabla^2 \widetilde{h}_\sigma(z)\|_\infty \\ &\le \int |\widetilde{h}(u)| \|\nabla^2 K_\sigma(z-u)\|_\infty du \\ &= \sigma^{-d-2} \int |\widetilde{h}(u)| \|\nabla^2 K((z-u)/\sigma)\|_\infty du \end{aligned}$$

$$\begin{aligned}&\leq \Lambda\sigma^{-d-2}\int \|z-u\|^{\alpha}\,\|\nabla^2 K((z-u)/\sigma)\|_{\infty} du\\&= \Lambda\sigma^{\alpha-2}\int \|u\|^{\alpha}\,\|\nabla^2 K(u)\|_{\infty} du \leq C\Lambda\sigma^{\alpha-2}\,,\end{aligned}$$

for some constant C depending only on K. This proves the second claim. □

3.B Omitted Proofs from Section 3.3

3.B.1 Proof of Lemma 23

Let $f : x \in B_{0,r} \mapsto \phi(x) - c_1\Lambda_r\|x\|^2$. Under condition (**H2**), recall that for all $R > 0$, $\|h\|_{\mathscr{C}^2(B_{0,R})} \leq \Lambda R^p$. Set

$$R = \sup\{\|x-y\| : x \in B_{0,r},\ y \in \partial^c\phi(B_{0,r})\},$$

and let $\Lambda_r = \Lambda R^p$. It then follows by the same argument as in the proof of Lemma 18 that the map

$$h_{\Lambda_r} : z \in B_{0,R} \mapsto h(z) - \frac{\Lambda_r}{2}\,\|z\|^2$$

is concave. Now, the assumptions on c in Lemma 17(iv) are satisfied under conditions (**H0**), (**H2**), and under the assumption of superlinearity of h, thus the assumption of local boundedness on ϕ ensures that $\partial^c\phi(x)$ is nonempty for all $x \in B_{0,r}$, and that ϕ admits the representation

$$\phi(x) = \inf_{y\in\partial^c\phi(B_{0,r})}\Big\{c(x,y) - \phi^c(y)\Big\}.$$

It follows that

$$\begin{aligned}f(x) &= \inf_{y\in\partial^c\phi(B_{0,r})}\{c(x,y)-\phi^c(y)\} - \frac{\Lambda_r}{2}\,\|x\|^2\\&= \inf_{y\in\partial^c\phi(B_{0,r})}\left\{h_{\Lambda_r}(x-y) + \frac{\Lambda_r}{2}\Big[\,\|y\|^2 - 2\langle x,y\rangle\Big] - \phi^c(y)\right\}.\end{aligned}$$

Notice that $\|x-y\| \leq R$ for all x, y appearing in the infimum of the final line in the above display, thus h_{Λ_r} is defined and concave therein. Similarly as in Lemma 18, the last line of the above display is thus an infimum of concave functions of x, implying that f is concave. To prove that f is Lipschitz, let $x \in B_{0,r}$ and choose a sequence $(y_k) \subseteq \mathcal{Y}$ such that

$$f(x) \geq c(x,y_k) - \phi^c(y_k) - \frac{\Lambda_r}{2}\,\|x\|^2 - k^{-1}.$$

Then, for all $x' \in B_{0,r}$ and $k \geq 1$,

$$\begin{aligned}f(x') - f(x) &\leq \left[c(x',y_k) - \phi^c(y_k) - \frac{\Lambda_r}{2}\|x'\|^2\right] - \left[c(x,y_k) - \phi^c(y_k) - \frac{\Lambda_r}{2}\,\|x\|^2\right] + k^{-1}\\&= h(x'-y_k) - h(x-y_k) - \frac{\Lambda_r}{2}\left[\|x'\|^2 - \|x\|^2\right] + k^{-1}\end{aligned}$$

$$\begin{aligned}
&\le \left(\sup_{z\in B_{0,R}} \|\nabla h(z)\|\right)\|x'-x\| + \frac{\Lambda_r}{2}(\|x'\|-\|x\|)(\|x'\|+\|x\|)+k^{-1}\\
&\le \left(\sup_{z\in B_{0,R}} \|\nabla h(z)\| + r\Lambda_r\right)\|x'-x\| + k^{-1}\\
&\le 2r\Lambda_r\|x'-x\|+k^{-1}.
\end{aligned}$$

The claim readily follows. □

3.B.2 Proof of Lemma 24

Part (i) is immediate by definition of c-conjugate. For part (ii), note that for any $x \in \operatorname{supp}(P_n)$,

$$\phi_n(x) = \inf_{y\in\mathbb{R}^d}\Big\{c(x,y)-\eta_n(y)\Big\} \ge \inf_{y\in\mathbb{R}^d}\Big\{c(x,y)-\Big[c(x,y)-f_n(x)\Big]\Big\} = f_n(x). \tag{3.41}$$

Similarly, for any $y \in \operatorname{supp}(Q_n)$, since $(f_n, g_n) \in \Phi_c(P_n, Q_n)$,

$$\begin{aligned}
\psi_n(y) &= \inf_{x\in\mathbb{R}^d}\Big\{c(x,y)-\phi_n(x)\Big\}\\
&\ge \inf_{x\in\mathbb{R}^d}\Big\{c(x,y)-\big[c(x,y)-\eta_n(y)\big]\Big\}\\
&= \eta_n(y)\\
&= \inf_{x\in\operatorname{supp}(P_n)}\Big\{c(x,y)-f_n(x)\Big\}\wedge \overline{R}_n\\
&\ge g_n(y)\wedge\overline{R}_n = g_n(y)\,,
\end{aligned}$$

where the final equality uses that g_n maps into $[0,\overline{R}_n]$. Since P_n and Q_n are finitely supported, either of the above inequalities is strict if and only if

$$\int \phi_n dP_n + \int \psi_n dQ_n > \int f_n dP_n + \int g_n dQ_n = \mathcal{T}_c(P_n,Q_n),$$

in violation of the optimality of (f_n, g_n). Therefore ϕ_n and ψ_n agree with f_n and g_n on $\operatorname{supp}(P_n)$ and $\operatorname{supp}(Q_n)$, and

$$\mathcal{T}_c(P_n,Q_n) = \int \phi_n dP_n + \int \psi_n Q_n. \tag{3.42}$$

To prove part (iii), note that η_n is nonnegative over $\mathbb{R}^d$, since f_n is nonpositive over $\operatorname{supp}(P_n)$. Therefore, for any $x\in\mathbb{R}^d$,

$$\phi_n(x) = \inf_{y\in\mathbb{R}^d}\Big\{c(x,y)-\eta_n(y)\Big\} \le h(0)-\eta_n(x)\le 0.$$

Furthermore, since η_n is bounded above by $\overline{R}_n$,

$$\phi_n(x) \ge \inf_{y\in\mathbb{R}^d}\Big\{c(x,y)-\overline{R}_n\Big\} \ge -\overline{R}_n, \tag{3.43}$$

Thus, $|\phi_n(x)| \leq \overline{R}_n$. Similarly, since ϕ_n is nonpositive, ψ_n is nonnegative, and for all $y \in \mathbb{R}^d$,

$$\psi_n(y) = \inf_{x \in \mathbb{R}^d} \Big\{ c(x,y) - \phi_n(x) \Big\} \leq h(0) - \phi_n(y) \leq \overline{R}_n,$$

where we used equation (3.43). Thus, $|\psi_n(x)| \leq \overline{R}_n$ as well.

Finally, to prove part (iv), equation (3.42) and the primal definition of $\mathcal{T}_c(P_n, Q_n) < \infty$ imply

$$\int \big[c(x,y) - \phi_n(x) - \psi_n(y) \big] d\pi_n(x,y) = 0.$$

By part (i), the integrand of the above display is nonnegative, thus

$$c(x,y) = \phi_n(x) + \psi_n(y), \quad \text{for all } (x,y) \in \operatorname{supp}(\pi_n).$$

Since ϕ_n and ψ_n are bounded by part (iii), it must then follow from Lemma 17(iv) that for any (x,y) satisfying the above display, $(x,y) \in \partial^c \phi_n(x)$ and $(y,x) \in \partial^c \psi_n(y)$. □

3.B.3 Proof of Lemma 25

Let $\mathcal{B}$ denote the set of all balls in $\mathbb{R}^d$. Recall that $\mathcal{B}$ has Vapnik-Chervonenkis dimension $d+2$, thus the Vapnik-Chervonenkis inequality (Vapnik and Chervonenkis, 1968) implies that for all $u > 0$,

$$\mathbb{P}\left(\sup_{B \in \mathcal{B}} |P_n(B) - P(B)| \geq u \right) \lesssim n^{d+2} \exp\left(-\frac{nu^2}{32} \right). \tag{3.44}$$

By the assumption of (γ, b)-super-Gaussianity, we have for all $\|y - x\| \leq 2$,

$$P(B_y) \gtrsim \int_{B_y} \exp\big(- \|u\|^2 / (2\gamma^2) \big) du \gtrsim \exp\big(- \|x\|^2 / \gamma^2 \big),$$

so that, for all $0 \leq j \leq J_n$,

$$\inf_{x \in I_j} \inf_{\|x-y\| \leq 2} P(B_y) \geq C_1 \exp(-\ell_j^2 / \gamma^2).$$

Thus setting $u = C_1 \exp(-\ell_j^2/\gamma^2)/2$ in equation (3.44) for all $0 \leq j \leq J_n$, and applying a union bound, leads to

$$\mathbb{P}(A_n^c) \lesssim n^{d+2} J_n \exp\left\{ -\frac{C_1^2}{128} n \exp(-2\ell_{J_n}^2/\gamma^2) \right\} = n^{d+2} J_n \exp\left\{ -\frac{C_1^2 \sqrt{n}}{128} \right\} \lesssim \frac{1}{n}. \tag{3.45}$$

The claim follows. □

3.B.4 Proof of Proposition 9

The proof proceeds using a similar argument as that of Proposition C.4 of Gangbo and McCann (1996). Under conditions (**H0**) and (**H3**), it follows from Lemma 17(iv) that $\partial^c\phi(x)$ is nonempty for all $x \in B_{r/2}$. For any $y \in \partial^c\phi(x)$, we have

$$\phi(x) = c(x,y) - \phi^c(y).$$

Let $v = x-y$. If $\|v\| \le r$ there is nothing to prove, so assume otherwise, and define $\xi = 1-\frac{r}{2\|v\|}$. Our assumption implies that $\xi \in [1/2, 1]$. Furthermore, define

$$u = x + (\xi - 1)v = x - \frac{r}{2}\left(\frac{v}{\|v\|}\right).$$

Then, the penultimate display leads to

$$h(v) - h(\xi v) = c(x,y) - c(u,y) = c(x,y) - \phi^c(y) - [c(u,y) - \phi^c(y)] \le \phi(x) - \phi(u) \le 2R.$$

This fact, together with the convexity and differentiability of h away from zero, under condition (**H3**), implies

$$\frac{r}{2}\langle \nabla h(\xi v), v/\|v\|\rangle \le 2R.$$

On the other hand, by condition (**H3**) we have $h(0) = 0$, thus by convexity of h,

$$\frac{h(\xi v)}{\|\xi v\|} \le \left\langle \nabla h(\xi v), \frac{\xi v}{\|\xi v\|}\right\rangle \le \frac{4R}{r}.$$

In particular, since $h(z) \gtrsim \kappa^{-1}\|z\|^p$ for all $\|z\| \ge 2$ under condition (**H3**), we have $\|\xi v\|^{p-1} \lesssim 4R/r$, thus since $\xi \ge 1/2$ and $r \ge 1$, $\|v\|^{p-1} \lesssim R$, and hence

$$\|y\|^{p-1} \lesssim \|x\|^{p-1} + R.$$

The claim follows. □

3.B.5 Proof of Lemma 26

Let $\bar{M}_{jk} = M_j$ and $\bar{U}_{jk} = U_j$ for all $j \ge 0$ and $k = 1, \dots, m_d$. Fix an enumeration $D_1, D_2, \dots$ (resp. $\bar{M}_1, \bar{M}_2, \dots$ and $\bar{U}_1, \bar{U}_2, \dots$) of the set $\{I_{jk} : j \ge 0, 1 \le k \le m_d\}$ (resp. $(\bar{M}_{jk})$, $(\bar{U}_{jk})$). Given a sequence $(a_j)_{j=1}^\infty$ of positive real numbers, let $p_j = N(\epsilon a_j, \mathcal{F}_{\bar{M}_j,\bar{U}_j}(D_j), L^\infty)$ and let $f_{j,1}, \dots, f_{j,p_j}$ be a ϵa_j-cover for $\mathcal{F}_{\bar{M}_j,\bar{U}_j}(D_j)$ in L^∞. By Lemma 20, we have

$$\log p_j \lesssim \left(\frac{\bar{U}_j + \mathrm{diam}(D_j)\bar{M}_j}{\epsilon a_j}\right)^{\frac{d}{2}}.$$

Now, it can be directly verified that the set

$$\left\{\sum_{j=1}^\infty f_{j,k_j} : k_j \in \{1, \dots, p_j\}, j \ge 0\right\}$$

forms an $\epsilon \left(\sum_{j=1}^\infty a_j^2 P_n(D_j)\right)^{1/2}$-cover of $\mathcal{K}_{M,U}$ in $L^2(P_n)$, which is of size $\prod_{j=1}^\infty p_j$. Thus,

$$\log N\left(\epsilon\left[\sum_{j=1}^\infty a_j^2 P_n(D_j)\right]^{1/2}, \mathcal{K}_{M,U}, L^2(P_n)\right) \le \sum_{j=1}^\infty \log p_j \lesssim \sum_{j=1}^\infty \left(\frac{\bar{U}_j + \mathrm{diam}(D_j)\bar{M}_j}{\epsilon a_j}\right)^{\frac{d}{2}}.$$

Now, set

$$a_j = \left(\bar{U}_j + \mathrm{diam}(D_j)\bar{M}_j\right)^{\frac{d}{d+4}} P_n(D_j)^{-\frac{2}{d+4}}.$$

Then

$$\sum_{j=1}^\infty \left(\frac{\bar{U}_j + \mathrm{diam}(D_j)\bar{M}_j}{a_j}\right)^{\frac{d}{2}} = \sum_{j=1}^\infty \left(\bar{U}_j + \mathrm{diam}(D_j)\bar{M}_j\right)^{\frac{2d}{d+4}} P_n(D_j)^{\frac{d}{d+4}},$$

and,

$$\sum_{j=1}^\infty a_j^2 P_n(D_j) \le \sum_{j=1}^\infty \left(\bar{U}_j + \mathrm{diam}(D_j)\bar{M}_j\right)^{\frac{2d}{d+4}} P_n(D_j)^{\frac{d}{d+4}}.$$

We deduce that for all $\epsilon > 0$,

$$\begin{aligned}\log N\left(\epsilon, \mathcal{K}_{M,U}, L^2(P_n)\right) &\lesssim \left(\sum_{j=1}^\infty a_j^2 P_n(D_j)\right)^{\frac{d}{4}} \sum_{j=1}^\infty \left(\frac{\bar{U}_j + \mathrm{diam}(D_j)\bar{M}_j}{\epsilon a_j}\right)^{\frac{d}{2}} \\ &\lesssim \left(\frac{1}{\epsilon}\right)^{\frac{d}{2}} \left(\sum_{j=1}^\infty \left(\bar{U}_j + \mathrm{diam}(D_j)\bar{M}_j\right)^{\frac{2d}{d+4}} P_n(D_j)^{\frac{d}{d+4}}\right)^{\frac{4+d}{4}}.\end{aligned}$$

The claim follows. □

3.B.6 Proof of Lemma 27

To prove the claim, it suffices to show that the variance of the supremum of the empirical process is of the order $(\log n)^{2r_4}/n$. By Boucheron, Lugosi, and Massart (2013, Theorem 11.1), it holds that

$$\begin{aligned}\mathrm{Var}\left[\sup_{f\in\mathcal{K}_{M,U}} \int f d(P_n - P)\right] &\le \frac{1}{n^2}\sum_{i=1}^n \mathbb{E}\left[\sup_{f\in\mathcal{K}_{M,U}} \left(f(X_i) - \mathbb{E}f(X_i)\right)^2\right] \\ &\lesssim \frac{1}{n}\mathbb{E}\left[\sup_{f\in\mathcal{K}_{M,U}} f^2(X_1)\right] \\ &= \frac{1}{n}\sum_{j=0}^\infty \int_{L_j}\left(\sup_{f\in\mathcal{K}_{M,U}} f^2(x)\right) dP(x)\end{aligned}$$

$$\begin{aligned}
&\le \frac{1}{n}\sum_{j=0}^{\infty} U_j^2 P(L_j) \\
&\lesssim \frac{1}{n}\sum_{j=0}^{\infty} (3^j \log n)^{2r_4} \exp(-c_1 3^{j\beta}) \\
&\lesssim \frac{(\log n)^{2r_4}}{n}.
\end{aligned}$$

The claim readily follows. □

3.B.7 Proof of Lemma 28

We begin with $\mathbb{E}|\Gamma_n|$. Since the quantity $\int \phi_0 d(P_n - P)$ remains unchanged if a constant is added to the map ϕ_0, there is no loss of generality in assuming $\phi_0(0) = 0$. By Theorem 10 and Lemma 23 it must then follow that $\phi_0(x) \lesssim 1 + \|x\|^q$ for a sufficiently large exponent $q \ge 1$, implying that

$$\mathbb{E}[\phi_0(X)^2] \lesssim 1 + \mathbb{E}[\|X\|^{2q}] \le C,$$

where the final inequality holds because P is (σ, β)-sub-Weibull, and thus admits $(2q)$-th moment bounded above by a constant depending only on q, σ and β, for all $q \ge 1$. Therefore, by Markov's inequality,

$$\mathbb{E}\left|\int \phi_0 d(P_n - P)\right| = \int_0^\infty \mathbb{P}\left(\left|\int \phi_0 d(P_n - P)\right| \ge u\right) du \le n^{-1/2} + \int_{n^{-\frac{1}{2}}}^\infty \frac{C}{nu^2} du \lesssim \frac{1}{\sqrt{n}}.$$

Applying a similar argument to ψ_0 leads to $\mathbb{E}|\Gamma_n| \lesssim n^{-1/2}$.

Turning to $\mathcal{X}_n$, notice that $|\xi_n(x)| \lesssim (\log n \, \|x\|)^{q'}$ for all $x \in \mathbb{R}^d$, for a sufficiently large constant $q' > 0$, thus it follows similarly as before that $\mathbb{E}[\mathcal{X}_n] \lesssim (\log n)^{q'} n^{-1/2} \lesssim n^{\epsilon - \frac{1}{2}}$ for any $\epsilon > 0$. □

3.B.8 Proof of Corollary 5

The claim is straightforward when $p \ge 2$. To prove the claim when $p \in (1, 2)$, abbreviate $h_p(x) = \|x\|^p$, and for all $\epsilon \in [0, 1]$ define the cost

$$h_{p,\epsilon}(x) = \left(\|x\|^2 + \epsilon^{\frac{2}{p}}\right)^{\frac{p}{2}} - \epsilon.$$

Lemma 29. *We have for all $p \in (1, 2)$ and all $\epsilon \in [0, 1]$,*

1. *$\|h_{p,\epsilon} - h_p\|_{L^\infty} \le 2\epsilon$.*
2. *$h_{p,\epsilon}$ satisfies condition **(H2)** with $\|h\|_{\mathscr{C}^2(B_{0,r})} \le \Lambda_\epsilon r^p$ for all $r \ge 1$, where $\Lambda_\epsilon = c_1 \epsilon^{1-\frac{2}{p}}$ for a universal constant $c_1 > 0$. Furthermore, $h_{p,\epsilon}$ satisfies condition **(H3)** with $\kappa = 2^{\frac{p}{2}-1}p$.*

Lemma 29(i) implies

$$\mathbb{E}\big|\mathcal{T}_{h_p}(P_n,Q_n)-\mathcal{T}_{h_p}(P,Q)\big| \leq \mathbb{E}\big|\mathcal{T}_{h_{p,\epsilon}}(P_n,Q_n)-\mathcal{T}_{h_{p,\epsilon}}(P,Q)\big|+4\epsilon,$$

which together with Lemma 29(ii) and Theorem 11 imply

$$\mathbb{E}\big|\mathcal{T}_{h_p}(P_n,Q_n)-\mathcal{T}_{h_p}(P,Q)\big| \lesssim \epsilon^{1-\frac{2}{p}}n^{-\frac{2}{d}}+\epsilon.$$

The right-hand side is minimized by choosing $\epsilon \asymp n^{-p/d}$, leading to the claim. It thus remains to prove Lemma 29.

3.B.8.1 Proof of Lemma 29

Notice that for all $x \in \mathbb{R}^d$,

$$|h_{p,\epsilon}(x)-h_p(x)| = \left|\Big(\,\|x\|^2+\epsilon^{\frac{2}{p}}\Big)^{\frac{p}{2}}-\epsilon-\|x\|^p\right| \leq \Big(\,\|x\|^2+\epsilon^{\frac{2}{p}}\Big)^{\frac{p}{2}}-\|x\|^p+\epsilon \leq 2\epsilon,$$

thus part (i) follows. To prove part (ii), choose the function $\omega(z)=(z^2+\epsilon^{2/p})^{p/2}-\epsilon$. We have $h(0)=0$, and for all $z>1$,

$$\omega'(z)=p(z^2+\epsilon^{2/p})^{\frac{p}{2}-1}z,$$

so that $h_{p,\epsilon}$ satisfies condition (**H3**) with $\kappa=2^{1-\frac{p}{2}}p$. It remains to prove the Hölder estimate. Clearly, $h_{p,\epsilon}\in\mathscr{C}^2_{\text{loc}}(\mathbb{R}^d)$ for all $\epsilon>0$, and

$$\nabla^2 h_{p,\epsilon}(x)=p(p-2)\Big(\,\|x\|^2+\epsilon^{\frac{2}{p}}\Big)^{\frac{p}{2}-2}xx^\top+p\Big(\,\|x\|^2+\epsilon^{\frac{2}{p}}\Big)^{\frac{p}{2}-1}I_d.$$

Therefore,

$$\big\|\nabla^2 h_{p,\epsilon}(x)\big\|_{\text{op}} \lesssim \Big(\,\|x\|^2+\epsilon^{\frac{2}{p}}\Big)^{\frac{p}{2}-2}\|x\|^2+\Big(\,\|x\|^2+\epsilon^{\frac{2}{p}}\Big)^{\frac{p}{2}-1} \lesssim \begin{cases}\epsilon^{1-\frac{2}{p}}, & \|x\|\leq\epsilon^{\frac{1}{p}},\\ \|x\|^{p-2}, & \|x\|>\epsilon^{\frac{1}{p}}.\end{cases}$$

We thus easily deduce that for all $r\geq 1$,

$$\|h_\epsilon\|_{\mathscr{C}^2(B_{0,r})}\lesssim r^p+\epsilon^{1-\frac{2}{p}}\leq\epsilon^{1-\frac{2}{p}}r^p,$$

and the claim follows. □

3.B.9 On the Super-Gaussianity Assumption

We close this Section with a simple characterization of super-Gaussianity which was stated in Section 3.3. Recall that we say a measure P is (γ,b)-super-Gaussian if $P(B_x)\geq b\cdot\mathbb{P}(Z\in B_x)$ for any $x\in\mathbb{R}^d$, where $Z\sim N(0,\gamma^2)$. Furthermore, we say that P admits a (γ_1,γ_2)-regular density (Polyanskiy and Wu, 2016) for some $\gamma_1,\gamma_2>0$ if P admits a density f with respect to the Lebesgue measure such that $\log f$ is differentiable and satisfies

$$\|\nabla\log f(x)\|\leq\gamma_1\,\|x\|+\gamma_2,\quad \text{for all } x\in\mathbb{R}^d.$$

Lemma 30. *Assume P admits a (γ_1,γ_2)-regular density. Then, there exist constants $\sigma,b>0$ such that P is (σ,b)-super-Gaussian.*

3.B.9.1 Proof of Lemma 30

By a first-order Taylor expansion, we have for all $x \in \mathbb{R}^d$,

$$|\log f(x) - \log f(0)| = |\nabla \log f(\widetilde{x})^\top x|,$$

for some $\|\widetilde{x}\| \leq \|x\|$. Therefore,

$$|\log f(x) - \log f(0)| \leq \|\nabla \log f(\widetilde{x})\| \|x\| \leq \gamma_1 \|x\|^2 + \gamma_2 \|x\|,$$

which entails

$$f(x) \geq \exp\Big\{ \log f(0) - \gamma_1 \|x\|^2 - \gamma_2 \|x\| \Big\} = f(0) \exp\Big\{ - \gamma_1 \|x\|^2 - \gamma_2 \|x\| \Big\}.$$

If $\|x\| \geq \gamma_2$, the above display is bounded below by $f(0)\exp(-(1+\gamma_1)\|x\|^2)$, while if $\|x\| \leq \gamma_2$, it is bounded below by $f(0)\exp(-\gamma_1 \|x\|^2 + \gamma_2^2)$. In either case, f is bounded below by a constant multiple of the $N(0, \gamma_1^{-1})$ density, thus the claim readily follows. □

3.C Omitted Proofs from Section 3.4

3.C.1 Proof of Theorem 12

Our proof of Theorem 12 follows similarly as that of Niles-Weed and Rigollet (2022), Theorem 11, which establishes a minimax lower bound for estimating p-Wasserstein distances. Our key extension of their proof technique is contained in the following result. Similarly as in the proof of Proposition 11, we write $T_0(z) = z + z_0$, where z_0 is defined in condition (**H4**).

Proposition 12. Assume the same conditions as Theorem 12. Given an integer $m \geq 1$, let u be the uniform distribution on $[m]$. Then, there exist universal constants $C_1, C_2 > 0$, a constant $C_{\lambda,\Lambda} > 0$ depending on λ, Λ, α, and a random function $F : [m] \to \mathcal{X}$, such that for any distribution q on $[m]$, we have

$$\begin{aligned} C_1 \lambda m^{-\alpha/d} \mathrm{TV}(q,u) - C_{\lambda,\Lambda}\sqrt{\frac{\chi^2(q,u)}{m}} &\leq \mathcal{T}_c\big(F_\# q, (T_0 \circ F)_\# u\big) - h(z_0) \\ &\leq C_2 \Lambda m^{-\alpha/d} \big(\chi^2(q,u)\big)^{\frac{\alpha}{d}} \mathrm{TV}(q,u)^{1-\frac{2\alpha}{d}} + C_{\lambda,\Lambda}\sqrt{\frac{\chi^2(q,u)}{m}}, \end{aligned}$$

with probability at least .9.

Proof. We prove the claim for $\alpha \in (1,2]$. An analogous argument may be used to prove the claim when $\alpha \in (0,1]$. Recall the notation of condition (**H4**). Similarly as in the proof of Proposition 11, there exists $\gamma > 0$ such that $\mathcal{X}_0 = B_{x_0,\gamma} \subseteq \mathcal{X}$ and such that $\mathcal{Y}_0 = T_0(\mathcal{X}_0) \subseteq \mathcal{Y}$, where $T_0(z) = z + z_0$. Now, it is a straightforward observation that $N(\epsilon, \mathcal{X}_0, \|\cdot\|) \geq c'\epsilon^{-d}$, for all $\epsilon \in (0,1)$ and for a constant $c' > 0$ depending only on d, γ, which implies that the ϵ-packing number of $\mathcal{X}$ under $\|\cdot\|$ is also greater than $c'\epsilon^{-d}$ (Wainwright (2019), Lemma 5.5). Therefore, there exists a set $\mathcal{G}_m = \{x_1, \dots, x_m\} \subseteq \mathcal{X}_0$ such that $\|x_i - x_j\| \gtrsim m^{-1/d}$ for all $i \neq j$. We let

F be selected uniformly at random from the set of bijections S_m from $[m] = \{1, \dots, m\}$ to $\mathcal{G}_m$.

We begin by proving the lower bound. Let $\widetilde{\pi}$ denote an optimal coupling between $F_\# q$ and $(T_0 \circ F)_\# u$, and let $\pi = (Id, T_0^{-1})_\# \widetilde{\pi} \in \Pi(F_\# q, F_\# u)$. We then have,

$$\begin{aligned}\mathcal{T}_c(F_\# q, (T_0 \circ F)_\# u) - h(z_0) &= \int \Big[h(y-x) - h(z_0) \Big] d\widetilde{\pi}(x,y) \\ &= \int \Big[h(y - x + z_0) - h(z_0) \Big] d\pi(x,y) \\ &\geq \int \langle \nabla h(z_0), y - x \rangle d\pi(x,y) + \lambda \int \|x-y\|^\alpha \, d\pi(x,y),\end{aligned} \tag{3.46}$$

by condition **(H4)**. We next bound the term $\Gamma_m = \int \langle \nabla h(z_0), y - x\rangle d\pi(x,y)$. Notice that

$$\begin{aligned}\mathbb{E}_F[\Gamma_m] &= \left\langle \nabla h(z_0), \frac{1}{m!} \sum_{G \in S_m} \sum_{j=1}^m (u(j) - q(j)) G(j) \right\rangle \\ &= \left\langle \nabla h(z_0), \sum_{j=1}^m (u(j) - q(j)) \left(\frac{1}{m!} \sum_{G \in S_m} G(j) \right) \right\rangle .\end{aligned}$$

The quantity $\frac{1}{m!} \sum_{G \in S_m} G(j)$ takes on the same value for all $j = 1, \dots, m$, thus we deduce from the above display that $\mathbb{E}_F[\Gamma_m] = 0$. Similarly, notice that for any $1 \leq j \neq k \leq m$,

$$\begin{aligned}\mathbb{E}_F \Big[\langle \nabla h(z_0), F(j) \rangle \langle \nabla h(z_0), F(k) \rangle \Big] &= \frac{1}{m!} \sum_{x \in \mathcal{G}_m} \sum_{\substack{y \in \mathcal{G}_m \\ y \neq x}} \sum_{\substack{G \in S_m \\ G(j) = x \\ G(k) = y}} \langle \nabla h(z_0), x \rangle \langle \nabla h(z_0), y \rangle \\ &= \frac{1}{m(m-1)} \sum_{x \neq y} \langle \nabla h(z_0), x \rangle \langle \nabla h(z_0), y \rangle =: M_{1,1},\end{aligned}$$

which is again constant in j, k. It follows that

$$\begin{aligned}&\mathbb{E}_F \left[\sum_{j \neq k} \langle \nabla h(z_0), F(j) \rangle \langle \nabla h(z_0), F(k) \rangle (q(j) - u(j))(q(k) - u(k)) \right] \\ &\quad = M_{1,1} \sum_{j \neq k} (q(j) - u(j))(q(k) - u(k)) = -M_{1,1} \sum_{j=1}^m \left(q(j) - \frac{1}{m} \right)^2 = -\frac{M_{1,1}}{m} \chi^2(q,u),\end{aligned}$$

whence, letting $M_2 := \mathbb{E}_F \Big[\langle \nabla h(z_0), F(j) \rangle^2 \Big]$, which itself is again constant in j, we obtain

$$\mathrm{Var}_F[\Gamma_n] = \mathbb{E}_F \left[\left(\sum_{j=1}^m \langle \nabla h(z_0), F(j) \rangle (q(j) - u(j)) \right)^2 \right]$$

$$= \sum_{j=1}^{m} \mathbb{E}_F \Big[\langle \nabla h(z_0), F(j) \rangle^2 \Big] (q(j) - u(j))^2 - \frac{M_{1,1}}{m} \chi^2(q,u) = \frac{M_2 - M_{1,1}}{m} \chi^2(q,u).$$

Therefore, by Markov's inequality, there exists a constant $C_{\lambda,\Lambda} > 0$ depending only on M_2, and hence only on λ, Λ, α, such that

$$\mathbb{P}\left(|\Gamma_n| \geq C_{\lambda,\Lambda} \sqrt{\frac{\chi^2(q,u)}{m}} \right) \leq .025. \tag{3.47}$$

Thus, returning to equation (3.46), and recalling that for all $x, y \in \mathcal{G}_m$, $\|x - y\| \gtrsim m^{-1/d} I(x \neq y)$, we deduce that for some $C_1 > 0$, with probability at least .975,

$$\begin{aligned} \mathcal{T}_c(F_\# q, (T_0 \circ F)_\# u) - h(z_0) &\geq C_1 \lambda m^{-\frac{\alpha}{d}} \mathbb{P}_\pi(X \neq Y) + \Gamma_m \\ &\geq C_1 \lambda m^{-\frac{\alpha}{d}} \mathrm{TV}(F_\# q, F_\# u) + \Gamma_m \\ &\geq C_1 \lambda m^{-\frac{\alpha}{d}} \mathrm{TV}(q,u) - C_{\lambda,\Lambda} \sqrt{\frac{\chi^2(q,u)}{m}}. \end{aligned} \tag{3.48}$$

We now prove the upper bound of the claim. Unlike before, we now let π denote an optimal coupling between $F_\# q$ and $F_\# u$, and $\widetilde{\pi} = (Id, T_0)_\# \pi \in \Pi(F_\# q, (T_0 \circ F)_\# u)$ a possibly suboptimal coupling. By assumption (**H1**), we then have

$$\begin{aligned} \mathcal{T}_c(F_\# q, (T_0 \circ F)_\# u) - h(z_0) &\leq \int \Big[h(y - x) - h(z_0) \Big] d\widetilde{\pi}(x,y) \\ &= \int \Big[h(y - x + z_0) - h(z_0) \Big] d\pi(x,y) \\ &\leq \int \langle \nabla h(z_0), y - x \rangle d\pi(x,y) + \Lambda \int \|x - y\|^\alpha \, d\pi(x,y) \\ &= \Gamma_m + \Lambda W_\alpha^\alpha(F_\# q, F_\# u). \end{aligned} \tag{3.49}$$

Now, by Niles-Weed and Rigollet (2022), Proposition 9, there exists a constant $C_2 > 0$ such that

$$W_\alpha^\alpha(F_\# u, F_\# q) \leq C_2 m^{-\alpha/d} \big(\chi^2(q,u) \big)^{\alpha/d} \mathrm{TV}(q,u)^{1 - \frac{2\alpha}{d}}.$$

After possibly modifying C_2, it follows from Markov's inequality that

$$\mathbb{P}\Big(W_\alpha^\alpha(F_\# u, F_\# q) \leq C_2 \Lambda m^{-\alpha/d} \big(\chi^2(q,u) \big)^{\alpha/d} \mathrm{TV}(q,u)^{1 - \frac{2\alpha}{d}} \Big) \geq .975,$$

so that, together with equations (3.47) and (3.49), we have with probability at least .95,

$$\mathcal{T}_c(F_\# u, (T_0 \circ F)_\# u) - h(z_0) \leq C_{\lambda,\Lambda} \sqrt{\frac{\chi^2(q,u)}{m}} + C_2 \Lambda m^{-\alpha/d} \big(\chi^2(q,u) \big)^{\alpha/d} \mathrm{TV}(q,u)^{1 - \frac{2\alpha}{d}}.$$

Combining this fact with equation (3.48) and a union bound leads to the claim. □

We now prove the main Theorem. In what follows, let $\mathcal{D}_m$ denote the set of probability distributions q on $[m]$ satisfying $\chi^2(q,u) \leq 9$. Also, given $\delta > 0$, let $\mathcal{D}_{m,\delta}^-$ denote the subset of distributions in $\mathcal{D}_m$ satisfying $\mathrm{TV}(q,u) \leq \delta$, and by $\mathcal{D}_m^+$ the subset of $\mathcal{D}_m$ satisfying $\mathrm{TV}(q,u) \geq 1/4$. Furthermore, set

$$\Delta_m = \frac{C_1 \lambda m^{-\alpha/d}}{16},$$

and $\delta = \left(\frac{C_1\lambda}{288\Lambda C_2}\right)^{\frac{1}{1-\frac{2\alpha}{d}}}$. Since $d \geq 5 > 2\alpha$, we may assume that m is large enough to satisfy

$$\Delta_m \geq 2C_{\lambda,\Lambda}\sqrt{\frac{9}{m}},$$

Then, by Proposition 12, for all $q \in \mathcal{D}_{m,\delta}^-$, we have with probability at least .9,

$$\begin{aligned}\mathcal{T}_c\big(F_\# q, (T_0 \circ F)_\# u\big) - h(z_0) &\leq C_2 \Lambda m^{-\alpha/d}\big(\chi^2(q,u)\big)^{\frac{\alpha}{d}} \mathrm{TV}(q,u)^{1-\frac{2\alpha}{d}} + C_{\lambda,\Lambda}\sqrt{\frac{\chi^2(q,u)}{m}} \\ &\leq \frac{C_1\lambda C_2 \Lambda m^{-\frac{\alpha}{d}} 9^{\frac{\alpha}{d}}}{288\Lambda C_2} + C_{\lambda,\Lambda}\sqrt{\frac{9}{m}} \leq \frac{\Delta_m}{2} + C_{\lambda,\Lambda}\sqrt{\frac{9}{m}} \leq \Delta_m.\end{aligned}$$

Similarly, for all $q \in \mathcal{D}_m^+$, we have with probability at least .9,

$$\begin{aligned}\mathcal{T}_c\big(F_\# q, (T_0 \circ F)_\# u\big) - h(z_0) &\geq C_1 \lambda m^{-\alpha/d}\mathrm{TV}(q,u) - C_{\lambda,\Lambda}\sqrt{\frac{\chi^2(q,u)}{m}} \\ &\geq \frac{C_1}{4}\lambda m^{-\alpha/d} - C_{\lambda,\Lambda}\sqrt{\frac{9}{m}} \geq 4\Delta_m - C_{\lambda,\Lambda}\sqrt{\frac{9}{m}} \geq 3\Delta_m.\end{aligned}$$

Now, for any given estimator $\widehat{\mathcal{T}}_n$ based on the independent samples $X_1, \dots, X_n$ and $Y_1, \dots, Y_n$, define the event $A = \{|\widehat{\mathcal{T}}_n - \mathcal{T}_c(F_\# q, (T_0 \circ F)_\# u)| \geq \Delta_m\}$. We have by Markov's inequality,

$$\begin{aligned}&\sup_{\substack{P \in \mathcal{P}(\mathcal{X}) \\ Q \in \mathcal{P}(\mathcal{Y})}} \mathbb{E}_{P,Q}\big|\widehat{\mathcal{T}}_n - \mathcal{T}_c(P,Q)\big| \\ &\qquad \geq \Delta_m \sup_{\substack{P \in \mathcal{P}(\mathcal{X}) \\ Q \in \mathcal{P}(\mathcal{Y})}} \mathbb{P}_{P,Q}\Big(|\widehat{\mathcal{T}}_n - \mathcal{T}_c(P,Q)| \geq \Delta_m\Big) \\ &\qquad \geq \frac{\Delta_m}{2}\left\{\sup_{q \in \mathcal{D}_{m,\delta}^-} \mathbb{E}_F \mathbb{P}_{F_\# q, (T_0\circ F)_\# u}[A] + \sup_{q \in \mathcal{D}_m^+} \mathbb{E}_F \mathbb{P}_{F_\# q, (T_0\circ F)_\# u}[A]\right\}. \qquad (3.50)\end{aligned}$$

Notice that for all $q \in \mathcal{D}_{m,\delta}^-$, we have

$$\begin{aligned}&\mathbb{E}_F \mathbb{P}_{F_\# q, (T_0\circ F)_\# u}[A] \\ &\quad \geq \mathbb{E}_F \mathbb{P}_{F_\# q, (T_0\circ F)_\# u}\Big(\widehat{\mathcal{T}}_n \geq h(z_0) + 2\Delta_m \text{ and } \mathcal{T}_c(F_\# q, (T_0 \circ F)_\# u) \leq h(z_0) + \Delta_m\Big)\end{aligned}$$

$$\begin{aligned}
&\geq \mathbb{E}_F \mathbb{P}_{F_\# q, (T_0 \circ F)_\# u}\Big(\widehat{\mathcal{T}}_n \geq h(z_0) + 2\Delta_m\Big) - \mathbb{P}_F\Big(\mathcal{T}_c(F_\# q, (T_0 \circ F)_\# u) > h(z_0) + \Delta_m\Big)\\
&\geq \mathbb{E}_F \mathbb{P}_{F_\# q, (T_0 \circ F)_\# u}\Big(\widehat{\mathcal{T}}_n \geq h(z_0) + 2\Delta_m\Big) - .1.
\end{aligned}$$

Similarly, for all $q \in \mathcal{D}_m^+$,

$$\begin{aligned}
&\mathbb{E}_F \mathbb{P}_{F_\# q, (T_0 \circ F)_\# u}[A]\\
&\quad\geq \mathbb{E}_F \mathbb{P}_{F_\# q, (T_0 \circ F)_\# u}\Big(\widehat{\mathcal{T}}_n \leq h(z_0) + 2\Delta_m \text{ and } \mathcal{T}_c(F_\# q, (T_v \circ F)_\# u) \geq h(z_0) + 3\Delta_m\Big)\\
&\quad\geq \mathbb{E}_F \mathbb{P}_{F_\# q, (T_0 \circ F)_\# u}\Big(\widehat{\mathcal{T}}_n \leq h(z_0) + 2\Delta_m\Big) - .1.
\end{aligned}$$

Returning to equation (3.50), we thus have,

$$\inf_{\widehat{\mathcal{T}}_n} \sup_{\substack{P \in \mathcal{P}(\mathcal{X}) \\ Q \in \mathcal{P}(\mathcal{Y})}} \mathbb{E}_{P,Q}\big|\widehat{\mathcal{T}}_n - \mathcal{T}_c(P, Q)\big| \geq \frac{\Delta_m}{2} \inf_{\psi} \left\{ \sup_{q \in \mathcal{D}_{m,\delta}^-} \mathbb{P}_q(\psi = 1) + \sup_{q \in \mathcal{D}_m^+} \mathbb{P}_q(\psi = 0) - 0.2 \right\},$$

where the infimum is over all tests based on the samples $X_1, \dots, X_n$ and $Y_1, \dots, Y_n$. By Proposition 10 of Niles-Weed and Rigollet (2022), the infimum on the right-hand side of the above display is bounded below by a constant if $m \asymp n \log n$. The claim then follows by definition of Δ_m. □

Chapter 4

Sequential Estimation of Optimal Transport Costs

4.1 Introduction

We saw in the previous chapter that the risk of the empirical optimal transport cost $\mathcal{T}_c(P_n, Q_n)$ typically degrades exponentially with the dimension. Implicit in our proofs, however, is the fact that this rate is driven entirely by the *bias* of the estimator: it is a simple observation that the *fluctuations* of the empirical optimal transport cost typically scale at the parametric rate, which is dimension-free. In fact, it can be shown using McDiarmid's inequality that, for bounded cost functions c, the empirical optimal transport cost is sub-Gaussian with parameter $O(1/n)$: there is a constant $C > 0$ such that for all $\delta \in (0,1)$,

$$\mathbb{P}\left(\mathcal{T}_c(P_n, Q_n) - \mathbb{E}\mathcal{T}_c(P_n, Q_n) \geq C\sqrt{\frac{\log(1/\delta)}{n}}\right) \leq \delta.$$

One of the main goals of this chapter will be to derive a *sequential* analogue of the above concentration inequality. For example, we will see in Corollary 12 below that the following time-uniform inequality holds for bounded costs c,

$$s\mathbb{P}\left(\forall n \geq 1 : \mathcal{T}_c(P_n, Q_n) - \mathbb{E}\mathcal{T}_c(P_{\bar{n}}, Q_{\bar{n}}) \geq C'\sqrt{\frac{\log(1/\delta) + \log\log n}{n}}\right) \leq \delta, \tag{4.1}$$

where $\bar{n} = \lceil n/2 \rceil$, and $C' > 0$ depends only on $\|c\|_\infty$. Bounds of the above type are sometimes known as *finite law of the iterated logarithm* bounds (Jamieson et al., 2014), and as we shall see in this chapter, they can be used to derive so-called *confidence sequences* for the population optimal transport cost $\mathcal{T}_c(P, Q)$ whenever the bias of the empirical Wasserstein distance is of the order $O(n^{-1/2})$.

It turns out that the techniques needed to obtain inequality (4.1) rely on a single property of optimal transport costs: their *convexity*, when viewed as functionals over the space of probability measures. Convexity is a property satisfied by many other natural functionals, such as most commonly-used divergences over the space of probability measures (cf. Section 4.2.1 below). Therefore, we will take a broader perspective in this chapter, and provide a general toolbox for performing rigorous *sequential* uncertainty quantification for convex functionals.

Problem Setting Let $\mathcal{X} \subseteq \mathbb{R}^d$. Throughout the sequel, we denote by $\Psi : \mathcal{P}(\mathcal{X}) \times \mathcal{P}(\mathcal{X}) \to \mathbb{R}_+$ a generic convex functional, that is, a functional satisfying the following condition: for all measures $\mu_1, \mu_2, \nu_1, \nu_2 \in \mathcal{P}(\mathcal{X})$ and all $\lambda \in [0, 1]$,

$$\Psi\big(\lambda\mu_1 + (1-\lambda)\mu_2, \lambda\nu_1 + (1-\lambda)\nu_2\big) \leq \lambda\Psi(\mu_1, \nu_1) + (1-\lambda)\Psi(\mu_2, \nu_2). \tag{4.2}$$

Given two independent sequences $(X_t)_{t=1}^\infty$ and $(Y_s)_{s=1}^\infty$ of i.i.d. observations arising respectively from unknown distributions $P, Q \in \mathcal{P}(\mathcal{X})$, we aim to construct a sequence of confidence intervals $(C_{ts})_{t,s=1}^\infty$ with the uniform coverage property

$$\mathbb{P}\big(\forall t, s \geq 1 : \Psi(P, Q) \in C_{ts}\big) \geq 1 - \delta, \tag{4.3}$$

for some pre-specified level $\delta \in (0, 1)$. Such a sequence $(C_{ts})_{t,s=1}^\infty$ is called a *confidence sequence*, and differs from the standard notion of confidence interval through the uniformity in times t, s of the probability in equation (4.3). As stated precisely in Section 4.3.3, the guarantee (4.3) is equivalent to the requirement that for all stopping times (τ, σ),

$$\mathbb{P}(\Psi(P, Q) \in C_{\tau\sigma}) \geq 1 - \delta. \tag{4.4}$$

The scope of potential applications of such confidence sequences is far-reaching. For instance, if Ψ is a metric between probability measures, then a confidence sequence C_{ts} directly gives rise to a sequential two-sample test for the null hypothesis $H_0 : P = Q$, where the null is rejected when $0 \notin C_{ts}$. Fixed-time, two-sample testing is a classical problem which continues to receive a wealth of attention, but *sequential, nonparametric* two-sample testing is relatively less explored; two exceptions include Balsubramani and Ramdas (2016), Lhéritier and Cazals (2018). We also note that confidence sequences for divergences can sometimes lead to confidence sequences for other estimands of interest, as illustrated in Section 4.4.5. Though our work is motivated by such practical applications, our results can also be viewed from a purely theoretical standpoint as deriving concentration inequalities for divergences, or other functionals, which hold uniformly over time. We are not aware of analogous time-uniform concentration inequalities in the literature for the majority of functionals studied explicitly in this chapter.

Our Contributions The primary contribution of our work is to provide a general recipe for deriving confidence sequences for convex functionals Ψ. Our key observation is that the process

$$M_{ts} = \Psi(P_t, Q_s) - \Psi(P, Q), \quad t, s \geq 1, \tag{4.5}$$

is a partially-ordered reverse submartingale, with respect to the so-called exchangeable filtration introduced below. Here, $P_t = (1/t)\sum_{i=1}^t \delta_{X_i}$ and $Q_s = (1/s)\sum_{i=1}^s \delta_{Y_j}$ denote empirical measures. A related property was previously identified by Pollard (1981) for suprema of empirical processes. This reverse submartingale property allows us to apply maximal inequalities to (functions of) $(M_{ts})_{t,s=1}^{\infty}$, which will lead to confidence sequences for $\Psi(P,Q)$ based on the plugin estimator $\Psi(P_t, Q_s)$. We note that this estimator is inconsistent for functionals requiring the absolute continuity of the probability measures being compared, such as φ-divergences for distributions supported over $\mathbb{R}^d$. We therefore extend our results by showing that the process in equation (4.5) continues to be a reverse submartingale in each of its indices when P_t and Q_s are replaced by their smoothed counterparts, $P_t \star \mathcal{K}_\sigma$ and $Q_s \star \mathcal{K}_\sigma$, where $\mathcal{K}_\sigma$ denotes a kernel with bandwidth σ.

We illustrate these findings by deriving explicit confidence sequences for optimal transport costs, the kernel Maximum Mean Discrepancy (MMD), Total Variation distance, and Kullback-Leibler divergence, among others, some for distributions over finite alphabets and others for arbitrary distributions. In all cases, we take care to track the effect of dimensionality, matching the best known rates is non-sequential settings. To the best of our knowledge, there are no other existing confidence sequences for these quantities, apart from the linear-time kernel MMD (Balsubramani and Ramdas, 2016). We also derive a sequential analogue of the celebrated Dvoretzky-Kiefer-Wolfowitz inequality (Dvoretzky, Kiefer, and Wolfowitz, 1956; Massart, 1990) quite differently from both Howard and Ramdas (2022) and a recent preprint by Maillard (2021), and demonstrate how these results can be used to obtain confidence sequences for convex functionals which do not necessarily arise from divergences.

4.2 Background

4.2.1 IPMs, Optimal Transport Costs, and φ-Divergences

In this section, we provide several examples of convex divergences, which form the main examples of functionals Ψ that we will appear in our development.

Let $\mathcal{X} \subseteq \mathbb{R}^d$ and let $\mathcal{A} \subseteq \mathcal{P}(\mathcal{X}) \times \mathcal{P}(\mathcal{X})$ be a set of pairs of probability measures. Throughout this paper, we use the term divergence to refer to any map $D : \mathcal{A} \to \mathbb{R}$ which is nonnegative and satisfies $D(P\|Q) = 0$ if $P = Q$, for all $(P,Q) \in \mathcal{A}$. When the divergence D is convex, we extend the domain of D from $\mathcal{A}$ to the entire set $\mathcal{P}(\mathcal{X}) \times \mathcal{P}(\mathcal{X})$ by letting D take the value ∞ wherever it is not defined—as such, convex divergences will always be understood as maps $D : \mathcal{P}(\mathcal{X}) \times \mathcal{P}(\mathcal{X}) \to \overline{\mathbb{R}}_+ := \mathbb{R}_+ \cup \{\infty\}$.

Optimal transport costs $\mathcal{T}_c$ form one example of convex divergences, which of course are of primary interest to us. The following are two other broad classes of convex divergences.

- **Integral Probability Metrics (IPMs).** Let $\mathcal{J}$ denote a set of Borel-measurable, real-valued functions on $\mathcal{X}$. The IPM (Müller, 1997) associated with $\mathcal{J}$ is given by

$$D_{\mathcal{J}}(P\|Q) = \sup_{f \in \mathcal{J}} \int f d(P-Q). \tag{4.6}$$

  For instance, when P and Q have supports contained in $\mathbb{R}$, the class of indicator functions $\mathcal{J} = \{I_{(-\infty,x]} : x \in \mathbb{R}\}$ gives rise to the Kolmogorov-Smirnov distance, $D_{\mathcal{J}}(P\|Q) = \|F-G\|_\infty$. When $\mathcal{J}$ is the unit ball of a reproducing kernel Hilbert space, $D_{\mathcal{J}}$ is called the (kernel) Maximum Mean Discrepancy (Gretton et al. (2012); see also Section 4.4.2). When $\mathcal{J}$ is the set of Borel-measurable maps $f : \mathbb{R}^d \to \mathbb{R}$ satisfying $\|f\|_\infty \leq 1$, $D_{\mathcal{J}}$ becomes the total variation distance $\|\cdot - \cdot\|_{\mathrm{TV}}$. When the function space $\mathcal{J}$ is sufficiently small, $D_{\mathcal{J}}(P_t\|Q_s)$ is a consistent estimator of $D_{\mathcal{J}}(P\|Q)$, and will form the basis of our confidence sequences. We refer to Sriperumbudur et al. (2012) for a study of convergence rates for such plugin estimators.

- **φ-Divergences.** Let $\varphi : \mathbb{R} \to \mathbb{R}$ be a convex function, and let $\nu \in \mathcal{P}(\mathcal{X})$ be a σ-finite measure which dominates both P and Q (for instance, $\nu = (P+Q)/2$). Let $p = dP/d\nu$ and $q = dQ/d\nu$ be the respective densities. Then, the φ-divergence (Ali and Silvey, 1966) between P and Q is given by

$$D_\varphi(P\|Q) = \int_{q>0} \varphi\left(\frac{p}{q}\right) dQ + P(q=0) \lim_{x\to\infty} \frac{\varphi(x)}{x}, \tag{4.7}$$

  with the convention that the second term of the above display is equal to zero whenever $P(q=0) = 0$, which is in particular the case if $P \ll Q$. For instance, assuming the latter condition holds, the Kullback-Leibler divergence $\mathrm{KL}(P\|Q) = \int \log\left(\frac{dP}{dQ}\right) dP$ corresponds to the map $\varphi(x) = x\log x$. The total variation distance is the unique nontrivial IPM which is also a φ-divergence (Sriperumbudur et al., 2012). φ-divergences also admit the variational representation

$$D_\varphi(P\|Q) \geq \sup_{g\in\mathcal{J}} \int \big[g dQ - (\varphi^* \circ g) dP\big], \tag{4.8}$$

  for any collection $\mathcal{J}$ of functions mapping $\mathcal{X}$ to $\mathbb{R}$. Equality holds in the above display if and only if the subdifferential $\partial\varphi(dQ/dP)$ contains an element of $\mathcal{J}$ (Nguyen, Wainwright, and Jordan, 2010).

  Unlike most common IPMs and optimal transport costs, φ-divergences are typically uninformative when P is not absolutely continuous with respect to Q, as the expression (4.7) becomes dominated by its second term. This fact sometimes prohibits the estimation of φ-divergences via the plugin estimator $D(P_t\|Q_s)$—for instance, P_t is almost surely not absolutely continuous with respect to Q_s when P and Q are both absolutely continuous with respect to the Lebesgue measure. One exception is the situation where P and Q are supported on countable sets, in which case Berend and Kontorovich (2013), Agrawal and Horel (2020), Guo and Richardson (2020), Cohen, Kontorovich, and Wolfer (2020), Han, Jiao, and Weissman (2015), Kamath et al. (2015), study concentration and convergence rates of the empirical measure P_t under the Kullback-Leibler and Total Variation divergences. We develop time-uniform bounds which build upon these results in Section 4.4.4. For distributions P and Q which are not countably-supported, distinct estimators have been developed by Nguyen, Wainwright, and Jordan (2010), Póczos and Schneider (2012), Sricharan, Raich, and Hero III (2010), Krishnamurthy et al. (2015), Rubenstein et al. (2019), Singh and Póczos (2014), Wang, Kulkarni, and Verdú (2005), Berrett and Samworth (2023), and references therein.

The following Lemma is standard, and stated without proof.

Lemma 31. *For any class $\mathcal{J}$ of Borel-measurable functions from $\mathbb{R}^d$ to $\mathbb{R}$, the IPM generated by $\mathcal{J}$ is convex. Furthermore, the φ-divergence generated by any convex function $\varphi : \mathbb{R} \to \mathbb{R}$ is convex. Finally, for any nonnegative cost function c, the optimal transport cost $\mathcal{T}_c$ is convex.*

Though our main focus is on the above divergences, our results also apply to one-sample convex functionals $\Phi : \mathcal{P}(\mathcal{X}) \to \mathbb{R}$, which satisfy $\Phi(\lambda\mu_1+(1-\lambda)\mu_2) \leq \lambda\Phi(\mu_1)+(1-\lambda)\Phi(\mu_2)$ for all $\lambda \in [0,1]$ and $\mu_1, \mu_2 \in \mathcal{P}(\mathcal{X})$. Notice that for any fixed distribution $Q \in \mathcal{P}(\mathcal{X})$, the map $\Phi(P) = D(P\|Q)$ is a convex functional—additional examples include the negative differential entropy (cf. Section 4.4.5) and certain expectation functionals (cf. Section 4.4.2).

Maximal martingales inequalities form a key tool in the development of confidence sequences, thus we provide an overview in what follows. Let $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability space, over which all processes hereafter will be taken. Before discussing time-reversed concepts, it is useful to first briefly overview standard martingales and filtrations.

4.2.2 Forward Filtrations, Martingales and Maximal Inequalities

A forward filtration is a sequence of σ-algebras $(\mathcal{F}_t)_{t=1}^\infty$ contained in $\mathcal{F}$ which is nondecreasing:

$$\mathcal{F}_t \subseteq \mathcal{F}_{t+1}, \quad \text{for all } t \geq 1.$$

Any process $(S_t)_{t=1}^\infty$ on Ω is adapted to its canonical (forward) filtration $(\mathcal{C}_t)_{t=1}^\infty$ defined by $\mathcal{C}_t = \sigma(S_1, \dots, S_t)$, for all $t \geq 1$. A (forward) martingale with respect to a (forward) filtration $(\mathcal{F}_t)_{t=1}^\infty$ is a stochastic process $(S_t)_{t=1}^\infty$ such that for all $t \geq 1$, S_t is $\mathbb{P}$-integrable, $\mathcal{F}_t$-measurable, and satisfies

$$\mathbb{E}[S_{t+1}|\mathcal{F}_t] = S_t, \quad \text{for all } t \geq 1.$$

Supermartingales and submartingales are respectively defined by replacing the equality in the above display by $\leq$ and $\geq$. When it causes no confusion, we frequently abbreviate $(S_t)_{t=1}^\infty$ and $(\mathcal{F}_t)_{t=1}^\infty$ by (S_t) and $(\mathcal{F}_t)$. The construction of confidence sequences typically relies on maximal inequalities. A prominent example is Ville's inequality (Ville, 1939), which states that any nonnegative supermartingale (S_t) satisfies

$$\mathbb{P}(\exists t \geq t_0 : S_t \geq u) \leq \frac{\mathbb{E}[S_{t_0}]}{u}, \quad \text{for all } u > 0 \text{ and all integers } t_0 \geq 1. \tag{4.9}$$

Inequality (4.9) is a time-uniform extension of Markov's inequality for nonnegative random variables. The unbounded range in Ville's inequality is made possible by nonnegativity, and the fact that supermartingales have nonincreasing expectations—see for instance Howard et al. (2020) for a formal proof. In contrast, submartingales admit nondecreasing expectations, and do not generally satisfy an infinite-horizon inequality such as (4.9). Nonnegative submartingales $(S_t)_{t=1}^\infty$ instead satisfy Doob's submartingale inequality (e.g. Durrett (2019), Theorem 4.4.2):

$$\mathbb{P}(\exists t \leq T : S_t \geq u) \leq \frac{\mathbb{E}[S_T]}{u} \quad \text{for all } u > 0 \text{ and any integer } T \geq 1. \tag{4.10}$$

The prototypical example of a martingale is a sum $S_t = \sum_{i=1}^t X_i$ of i.i.d. random variables $(X_t)_{t=1}^\infty \subseteq \mathcal{X} \subseteq \mathbb{R}^d$, with respect to its canonical filtration $\mathcal{C}_t = \sigma(X_1, \dots, X_t)$. As described in Section 4.2.5, a wealth of existing works have developed confidence sequences for the expected value of a sequence of i.i.d. random variables, on the basis of inequalities such as (4.9) and (4.10).

4.2.3 Reverse Filtrations, Martingales and Maximal Inequalities

An alternate approach is to apply maximal inequalities to the sample mean itself, namely to the process $R_t = (1/t)\sum_{i=1}^t X_i$. This approach is relatively underexplored but turns out to be well-suited to our goals. Unlike (S_t), the process (R_t) is a *reverse* martingale. To elaborate, a reverse filtration is a *nonincreasing* sequence of σ-algebras $(\mathcal{F}_t)_{t=1}^\infty$ contained in $\mathcal{F}$:

$$\mathcal{F}_t \supseteq \mathcal{F}_{t+1} \quad \text{for all } t \geq 1.$$

A $\mathbb{P}$-integrable process $(R_t)_{t=1}^\infty$ is then said to be a reverse martingale with respect to $(\mathcal{F}_t)_{t=1}^\infty$ if for all $t \geq 1$, R_t is $\mathcal{F}_t$-measurable and

$$\mathbb{E}[R_t|\mathcal{F}_{t+1}] = R_{t+1}, \quad \text{for all } t \geq 1.$$

Reverse submartingales and supermartingales are defined analogously, with the above equality respectively replaced by $\geq$ and $\leq$. The sample average $R_t = (1/t)\sum_{i=1}^t X_i$ defines a reverse martingale with respect to its canonical reverse filtration $\big(\sigma(R_t, R_{t+1}, \dots)\big)_{t=1}^\infty$. (R_t) is also a reverse martingale with respect to a richer filtration known as the *symmetric* or *exchangeable filtration* (Pollard, 2002; Durrett, 2019).

Definition 1 (Exchangeable Filtration)**.** Given a sequence of random variables $(X_t)_{t=1}^\infty$, the exchangeable filtration is the reverse filtration $(\mathcal{E}_t)_{t=1}^\infty$, where $\mathcal{E}_t$ denotes the σ-algebra generated by all real-valued Borel-measurable functions of $X_1, X_2, \dots$ which are permutation-symmetric in their first t arguments.

Given a sequence $(X_t)_{t=1}^\infty$ of exchangeable random variables, recall that $P_t = (1/t)\sum_{i=1}^t \delta_{X_i}$ denotes their empirical measure. It can be seen that the process $(P_t(B))_{t\geq 1}$ is a reverse martingale with respect to the exchangeable filtration $(\mathcal{E}_t)_{t=1}^\infty$, for any fixed Borel set $B \subseteq \mathbb{R}^d$. This property makes $(P_t)_{t=1}^\infty$ into a so-called measure-valued reverse martingale (Kallenberg, 2006). In fact, the converse nearly holds true: if (P_t) is a measure-valued reverse martingale, and if (X_t) is stationary, then (X_t) is exchangeable (Bladt and Shaiderman, 2023). Note that the exchangeability condition is weaker than the i.i.d. assumption which we shall assume for the majority of this paper. We will later see that, with some care, this measure-valued reverse martingale gets effectively translated into a (real-valued) reverse submartingale property for convex functionals.

A key technical tool for handling real-valued, reverse submartingales will be the following analogue of Ville's inequality (4.9), first proved by Doob (1940; Theorem 1.1, p. 458) for reverse martingales; see also Lee (1990; Theorem 3, p. 112).

Theorem 13 (Ville's Inequality for Nonnegative Reverse Submartingales)**.** *Let $(R_t)_{t=1}^\infty$ be a nonnegative reverse submartingale with respect to a reverse filtration $(\mathcal{F}_t)_{t=1}^\infty$. Then, for any* *integer $t_0 \geq 1$ and real number $u > 0$,*

$$\mathbb{P}\big(\exists t \geq t_0 : R_t \geq u\big) \leq \frac{\mathbb{E}[R_{t_0}]}{u}.$$

Since Theorem 13 plays a central role in our development, we provide two self-contained proofs in Section 4.B for completeness. As an example, it can be deduced from Theorem 13 that for any sequence of exchangeable random variables $(X_t)_{t=1}^{\infty}$,

$$\mathbb{P}\left(\sup_{t\geq 1}\left|\frac{1}{t}\sum_{i=1}^{t} X_i\right| \geq u\right) \leq \frac{\mathbb{E}|X_1|}{u}, \tag{4.11}$$

for all $u > 0$. This can be seen as a strengthening of Markov's inequality[1], whose left-hand side is $\mathbb{P}(|X_1| \geq u)$.

4.2.4 Partially Ordered Martingales

In order to handle the two-sample process (M_{ts}) in equation (4.5), we also employ (reverse) martingales indexed by $\mathbb{N}^2$. We endow $\mathbb{N}^2$ with the standard partial ordering, that is we write $(t, s) \leq (t', s')$ for all $(t, s), (t', s') \in \mathbb{N}^2$ such that $t \leq t'$ and $s \leq s'$. The notation $(t, s) < (t', s')$ indicates that at least one of the componentwise inequalities holds strictly. The definitions of filtration and martingale extend to this setting in the natural way: a forward (or reverse) filtration is a sequence $(\mathcal{F}_{ts})_{t,s\geq 1}$ of σ-algebras which is nondecreasing (or nonincreasing) with respect to the partial ordering, and a forward (or reverse) martingale $(S_{ts})_{t,s\geq 1}$ (or $(M_{ts})_{t,s\geq 1}$) is an $L^1(\mathbb{P})$ process adapted to $(\mathcal{F}_{ts})$ which satisfies $\mathbb{E}[S_{ts}|\mathcal{F}_{t's'}] = S_{t's'}$ for all $(t, s) > (t', s')$ (or $\mathbb{E}[R_{ts}|\mathcal{F}_{t's'}] = R_{t's'}$ for all $(t, s) < (t', s')$). When the latter equalities are replaced by the inequalities $\leq$ and $\geq$, (M_{ts}) is respectively called a (reverse) supermartingale and submartingale. We refer to Ivanoff and Merzbach (1999) for a survey.

Cairoli (1970) showed that a direct analogue of Ville's inequality cannot hold for partially ordered nonnegative martingales—the inequality $\mathbb{P}(\exists t, s \geq 1 : S_{ts} \geq u) \leq \mathbb{E}[S_{11}]/u$ does not generally hold for all $u > 0$. Distinct maximal inequalities for partially ordered forward and reverse martingales have nevertheless been established by Christofides and Serfling (1990) under suitable moment assumptions, and under the so-called conditional independence (CI) assumption introduced by Cairoli and Walsh (1975). A reverse filtration $(\mathcal{F}_{ts})_{t,s\geq 1}$ is said to satisfy the CI property if for all $t, t', s, s' \geq 1$,

$$\mathbb{E}\big\{\mathbb{E}[\,\cdot \mid \mathcal{F}_{ts'}] \bigm| \mathcal{F}_{t's}\big\} = \mathbb{E}\big\{\ \cdot\ \mid \mathcal{F}_{(t\vee t')(s\vee s')}\big\}, \tag{4.12}$$

or equivalently, that $\mathcal{F}_{ts'}$ and $\mathcal{F}_{t's}$ are conditionally independent given $\mathcal{F}_{(t\vee t')(s\vee s')}$ (Merzbach, 2003). The following Ville-type inequality for partially ordered reverse submartingales can be obtained by employing Corollary 2.9 of Christofides and Serfling (1990).

[1] In a separate line of work, we have built upon this observation to derive several other strictly sharper variants of Markov's inequality, and shown how they may be used to improve the power of a variety of batch and sequential nonparametric testing procedures (Ramdas and Manole, 2023).

Proposition 13. Let $(R_{ts})_{t,s\geq 1}$ be a nonnegative reverse submartingale with respect to a reverse filtration $(\mathcal{F}_{ts})_{t,s\geq 1}$ satisfying the conditional independence assumption. Assume that for some $\alpha > 1$, $R_{ts} \in L^\alpha(\mathbb{P})$ for all $t, s \geq 1$. Then, for all $u > 0$,

$$\mathbb{P}(\exists t \geq t_0, s \geq s_0 : R_{ts} \geq u) \leq \left(\frac{\alpha}{\alpha-1}\right)^\alpha \frac{\mathbb{E}[R_{t_0 s_0}^\alpha]}{u^\alpha}.$$

The special case $\alpha = 2$ was stated in Corollary 2.10 of Christofides and Serfling (1990). A proof and further discussion of this result is given in Section 4.B.2, and forms the basis for our two-sample results.

4.2.5 Time-Uniform Confidence Sequences

Confidence sequences are defined similarly as in (4.3) for estimands other than divergences. Given a functional $\theta \equiv \theta(P)$ of interest, and an error level $\delta \in (0,1)$, a $(1-\delta)$-confidence sequence $(C_t)_{t=1}^\infty$ based on an i.i.d. sequence of random variables $(X_t)_{t=1}^\infty$ from P is a sequence of sets $C_t \in \sigma(X_1, \dots, X_t)$ satisfying $\mathbb{P}(\exists t \geq 1 : \theta \notin C_t) \leq \delta$. When θ is real-valued, we say two sequences (ℓ_t) and (u_t) are lower and upper confidence sequences if $\mathbb{P}(\exists t \geq 1 : \theta \leq \ell_t) \leq \delta$ and $\mathbb{P}(\exists t \geq 1 : \theta \geq u_t) \leq \delta$ respectively. Confidence sequences were pioneered by Robbins, Darling, Siegmund and Lai (Darling and Robbins, 1967; Robbins and Siegmund, 1969; Lai, 1976), and new techniques have been recently developed that enable their extensions to new, nonparametric settings (Kaufmann and Koolen, 2021; Howard et al., 2021). This resurgence of interest in sequential analysis has been driven in part by its applications to best-arm identification algorithms for multi-armed bandits (Jamieson et al., 2014; Kaufmann, Cappé, and Garivier, 2016; Shin, Ramdas, and Rinaldo, 2021) and reinforcement learning (Karampatziakis, Mineiro, and Ramdas, 2021), to name a few.

The mean of a distribution is perhaps the target of inference which has received the most attention in prior work on confidence sequences. As described in Section 4.2.2, the process $S_t = \sum_{i=1}^t (X_i - \mu)$ forms a canonical example of a forward martingale, where μ denotes the common (finite) mean of the i.i.d. random variables X_i. To obtain a confidence sequence for μ, a maximal inequality such as that of Ville (see equation (4.9)) cannot be directly be applied to (S_t), however, since it is not a nonnegative process. Inspired by the Cramér-Chernoff method for deriving concentration inequalities, it is instead natural to consider the nonnegative process

$$U_t(\lambda) = \exp(\lambda S_t), \quad t \geq 1,$$

for $\lambda > 0$. Here, we assume a tail assumption is placed on X_1, such that it admits a finite cumulant generating function satisfying $\log\{\mathbb{E}[\exp(\lambda(X_1 - \mu))]\} \leq \phi(\lambda)$, for some (say, known) map $\phi : [0, \lambda_{\max}) \to \mathbb{R}$, where $\lambda_{\max} > 0$. The exponential process $(U_t(\lambda))$ is a nonnegative submartingale by Jensen's inequality, and forms the basis of several confidence sequences described below. A distinct line of work constructs confidence sequences for μ by downweighting this process to recover a (super)martingale. For instance, it can be verified that

$$L_t(\lambda) = \exp\{\lambda S_t - t\phi(\lambda)\}, \quad t \geq 1,$$

is a nonnegative supermartingale with respect to the canonical filtration. Variants of the process $(L_t(\lambda))$ appear in a long line of work aimed at deriving sequential concentration inequalities for means—see Howard et al. (2020) for a comprehensive review of such approaches.

Applying a maximal martingale inequality to either $(U_t(\lambda))$ or $(L_t(\lambda))$ does not, on its own, lead to satisfactory confidence sequences for μ. Infinite-horizon maximal inequalities are not available for submartingales $(U_t(\lambda))$, and even for the supermartingale $(L_t(\lambda))$, a direct application of Ville's inequality leads to a confidence sequence for μ with nonvanishing length. To obtain confidence sequences with lengths scaling at the optimal rate $O(\sqrt{\log\log t/t})$, implied by the law of the iterated logarithm, it is instead common to use a variant of the "method of mixtures", for example by repeatedly applying a maximal inequality over geometrically-spaced epochs in time—such methods are often known as peeling, chaining, or stitching. Jamieson et al. (2014); Zhao et al. (2016) use stitching arguments based on $(U_t(\lambda))$ and Doob's submartingale inequality, while Garivier (2013); Kaufmann, Cappé, and Garivier (2016); Howard et al. (2021), use $(L_t(\lambda))$ and Ville's inequality. The resulting confidence sequences decay at similar rates, though with varying constants and tail assumptions—see Howard et al. (2021); Waudby-Smith and Ramdas (2024) for a comparison of such approaches.

We have found that the above framework cannot be easily extended to generic functionals, of the type that we have in mind in this chapter. Our main results will instead hinge upon *reverse* submartingales—a rarely used tool for deriving confidence sequences.

4.3 Confidence Sequences for Convex Functionals

Let $(X_t)_{t=1}^\infty$ and $(Y_s)_{s=1}^\infty$ respectively denote independent sequences of i.i.d. observations from two distributions $P, Q \in \mathcal{P}(\mathcal{X})$, with support contained in a set $\mathcal{X} \subseteq \mathbb{R}^d$. Given convex functionals $\Phi : \mathcal{P}(\mathcal{X}) \to \overline{\mathbb{R}}$ and $\Psi : \mathcal{P}(\mathcal{X}) \times \mathcal{P}(\mathcal{X}) \to \overline{\mathbb{R}}$, the goal of this section is to derive confidence sequences for $\Phi(P)$ and $\Psi(P,Q)$. Our primary interest is in the special case where Ψ is a convex divergence D, and $\Phi = D(\cdot\|Q)$ when Q is known, but we formulate our results for arbitrary convex functionals in the interest of generality. We shall make use of the processes

$$N_t = \Phi(P_t) - \Phi(P), \quad M_{ts} = \Psi(P_t, Q_s) - \Psi(P,Q), \quad t, s \geq 1. \tag{4.13}$$

We prove in Section 4.3.1, that $(M_{ts})_{t,s\geq 1}$ and $(N_t)_{t\geq 1}$ are reverse submartingales with respect to suitable filtrations, whose choice is further discussed in Section 4.3.3. We then derive maximal inequalities for these processes, from which lower confidence sequences will follow using an epoch-based analysis. We follow a distinct strategy to obtain upper confidence sequences in Section 4.3.2.

4.3.1 Lower Confidence Sequences via Reverse Submartingales

Let $(\mathcal{E}_t^X)_{t=1}^\infty$ and $(\mathcal{E}_s^Y)_{s=1}^\infty$ denote the exchangeable filtrations associated with the sequences $(X_t)_{t=1}^\infty$ and $(Y_s)_{s=1}^\infty$ respectively, and define $\mathcal{E}_{ts} = \mathcal{E}_t^X \bigvee \mathcal{E}_s^Y$ for all $t, s \geq 1$. As stated in Section 4.2.2, the sequences of empirical measures (P_t) and (Q_s) form measure-valued reverse martingales with respect to $(\mathcal{E}_t^X)$ and $(\mathcal{E}_s^Y)$ respectively. This fact suggests that any convex functional evaluated at (P_t, Q_s) is a reverse submartingale, as we now show.

Theorem 14. *Let Φ, Ψ be convex functionals such that $\Phi(P_t), \Psi(P_t, Q_s) \in L^1(\mathbb{P})$ for all $t, s \geq 1$. Then,*

(i) $(\Phi(P_t))_{t\geq 1}$ is a reverse submartingale with respect to $(\mathcal{E}_t^X)$.

(ii) $(\Psi(P_t, Q_s))_{t,s\geq 1}$ is a partially ordered reverse submartingale with respect to $(\mathcal{E}_{ts})$.

Theorem 14(i) is known in the special case where Φ is the supremum of an empirical processes—see for instance Pollard (1981) and Lemma 2.4.5 of van der Vaart and Wellner (1996). Our proof below is inspired by these works, and in particular extends them to general convex functionals and to partially ordered filtrations.

Notice that the processes in Theorem 14 lie in $L^1(\mathbb{P})$, and are therefore assumed to be measurable. When this assumption is removed, it can be shown that there exist measurable *covers* of the processes in Theorem 14 for which the result continues to hold—this is the approach taken by van der Vaart and Wellner (1996) for suprema of empirical processes; see also Pollard (1981) and Strobl (1995). With this modification, we believe that all subsequent claims in this paper can be made to hold in outer probability, but we prefer to retain the measurability condition to avoid being overwhelmed by technicalities. In particular, (N_t) and (M_{ts}) are tacitly assumed to be measurable throughout the sequel.

Proof of Theorem 14. We will prove Theorem 14(ii) and a similar argument can be used to prove Theorem 14(i). For all $t, s \geq 1$ $\Psi(P_t, Q_s)$ is invariant to permutations of $X_1, \dots, X_t$ and of $Y_1, \dots, Y_s$. It follows that $(\Psi(P_t, Q_s))$ is adapted to $(\mathcal{E}_{ts})$, thus it suffices to prove the reverse submartingale property. Fix $t, s \geq 1$. Define the $(t+1)$ different "leave-one-out" empirical measures

$$P_t^i = \frac{1}{t}\left[\sum_{j=1}^{i-1} \delta_{X_j} + \sum_{j=i+1}^{t+1} \delta_{X_j}\right], \quad i = 1, 2, \dots, t+1.$$

Then $P_{t+1} = \frac{1}{t+1}\sum_{i=1}^{t+1} P_t^i$, and the convexity of Ψ implies

$$\Psi(P_{t+1}, Q_s) \leq \frac{1}{t+1}\sum_{i=1}^{t+1} \Psi(P_t^i, Q_s).$$

Since $\Psi(P_{t+1}, Q_s)$ is $\mathcal{E}_{(t+1)s}$-measurable, we deduce that

$$\Psi(P_{t+1}, Q_s) \leq \frac{1}{t+1}\sum_{i=1}^{t+1} \mathbb{E}[\Psi(P_t^i, Q_s)|\mathcal{E}_{(t+1)s}]. \tag{4.14}$$

Since $P_t^{t+1} = P_t$ by definition, the claim will follow upon proving the key identity

$$\mathbb{E}[\Psi(P_t^i, Q_s)|\mathcal{E}_{(t+1)s}] = \mathbb{E}[\Psi(P_t, Q_s)|\mathcal{E}_{(t+1)s}], \quad i = 1, \dots, t. \tag{4.15}$$

Notice that we can write $\mathcal{E}_{(t+1)s} = \mathcal{E}_{t+1}^X \bigvee \mathcal{E}_s^Y = \sigma(\mathcal{I})$ where

$$\mathcal{I} = \{A^X \cap A^Y : A^X \in \mathcal{E}_{t+1}^X, A^Y \in \mathcal{E}_s^Y\}.$$

$\mathcal{I}$ is clearly a π-system. To prove (4.15), it will thus suffice to prove that for any set $A = A^X \cap A^Y$, with $A^X \in \mathcal{E}_{t+1}^X$ and $A^Y \in \mathcal{E}_s^Y$, and for all $1 \leq i \leq t$, we have $\mathbb{E}[\Psi(P_t^i, Q_s)I_A] = \mathbb{E}[\Psi(P_t, Q_s)I_A]$, where $I_A : \Omega \to \{0, 1\}$ is the indicator function of A and $I_A = I_{A^X} I_{A^Y}$.

Notice that I_{A^X} is $\mathcal{E}_{t+1}^X$-measurable, thus it is a function f_{A^X} of $X_1, X_2, \dots$ which is permutation symmetric in $X_1, \dots, X_{t+1}$. For convenience, we write $I_{A^X} = f_{A^X}(X_1, X_2, \dots)$, whence we may also write $I_A = I_{A^Y} f_{A^X}(X_1, X_2, \dots)$. Now, let $\tau : \mathbb{N} \to \mathbb{N}$ be the permutation such that $\tau(j) = j$ if $j \notin \{t+1, i\}$ and $\tau(t+1) = i$, $\tau(i) = t+1$. By exchangeability of $X_1, X_2, \dots$, and by their independence from $Y_1, Y_2, \dots$, we have $(X_1, Y_1, X_2, Y_2, \dots) \overset{d}{=} (X_{\tau(1)}, Y_1, X_{\tau(2)}, Y_2, \dots)$, whence,

$$\begin{aligned}\mathbb{E}[\Psi(P_t, Q_s)I_A] &= \mathbb{E}[\Psi(P_t^{t+1}, Q_s) f_{A^X}(X_1, X_2, \dots) I_{A^Y}] \\ &= \mathbb{E}[\Psi(P_t^i, Q_s) f_{A^X}(X_{\tau(1)}, X_{\tau(2)}, \dots) I_{A^Y}].\end{aligned}$$

Since f_{A^X} is permutation-symmetric in its first $t+1$ arguments, and the permutation τ fixes all natural numbers greater or equal to $t+2$, we obtain

$$f_{A^X}(X_{\tau(1)}, X_{\tau(2)}, \dots,) = f_{A^X}(X_1, X_2, \dots),$$

implying that

$$\mathbb{E}[\Psi(P_t, Q_s) f_{A^X}(X_1, X_2, \dots) I_{A^Y}] = \mathbb{E}[\Psi(P_t^i, Q_s) f_{A^X}(X_1, X_2, \dots) I_{A^Y}].$$

Equation (4.15) now follows. Returning to equation (4.14) we deduce that

$$\Psi(P_{t+1}, Q_s) \leq \mathbb{E}[\Psi(P_t, Q_s) | \mathcal{E}_{(t+1)s}].$$

A symmetric argument shows that $\Psi(P_t, Q_{s+1}) \leq \mathbb{E}[\Psi(P_t, Q_s) | \mathcal{E}_{t(s+1)}]$, implying that $(\Psi(P_t, Q_s))$ is a partially ordered reverse submartingale with respect to $(\mathcal{E}_{ts})$. □

It is apparent from the proof that the convexity requirement is stronger than necessary. For example, the following more general statement can be inferred from the proof of Theorem 14. We record it formally as it may be of independent interest.

Proposition 14. Suppose $(R_t)_{t=1}^\infty$ is an $L^1(\mathbb{P})$ process of the form $R_t = f_t(X_1, \dots, X_t)$, for some sequence of permutation invariant maps $f_t : \mathcal{X}^t \to \mathbb{R}$. Suppose further that for all $t \geq 1$,

$$R_{t+1} \leq \frac{1}{t+1} \sum_{i=1}^{t+1} f_t(X_1, \dots, X_{i-1}, X_{i+1}, \dots, X_{t+1}). \tag{4.16}$$

Then, (R_t) is a reverse submartingale with respect to $(\mathcal{E}_t^X)$. Furthermore, if the above display holds with equality, (R_t) is a reverse martingale with respect to $(\mathcal{E}_t^X)$.

We refer to condition (4.16) as the *leave-one-out property*. In Sections 4.4.2 and 4.4.6 we will briefly make use of processes which cannot be expressed as evaluations of a convex functional at the empirical measure, but which satisfy the leave-one-out property.

Theorem 14 permits the use of maximal inequalities discussed in Sections 4.2.3 and 4.2.4 for reverse submartingales. In view of generalizing the Cramér-Chernoff technique (Boucheron, Lugosi, and Massart, 2013) to our sequential setup, we shall state our bounds under the assumption that N_t and M_{ts} admit finite cumulant generating functions over an interval $[0, \lambda_{\max})$, and are upper bounded by known convex functions $\psi_{ts}, \psi_t : [0, \lambda_{\max}) \to \mathbb{R}$,

$$\log\Big\{\mathbb{E}\big[\exp(\lambda M_{ts})\big]\Big\} \leq \psi_{ts}(\lambda), \quad \log\Big\{\mathbb{E}\big[\exp(\lambda N_t)\big]\Big\} \leq \psi_t(\lambda), \quad t, s \geq 1,\ \lambda \in [0, \lambda_{\max}). \tag{4.17}$$

We shall discuss in Section 4.4 how such tail assumptions may be replaced by tail assumptions on the distributions P and Q themselves, for various special cases of functionals. Under equation (4.17), we obtain the following result.

Proposition 15. Let Φ, Ψ be convex functionals such that the processes (M_{ts}) and (N_t) satisfy the bounds of equation (4.17). Then, for all $u > 0$, and all integers $t_0, s_0 \geq 1$, the following hold.

(i) (One-Sample) $\mathbb{P}(\exists t \geq t_0 : N_t \geq u) \leq \exp(-\psi_{t_0}^*(u))$.

(ii) (Two-Sample) $\mathbb{P}(\exists t \geq t_0, s \geq s_0 : M_{ts} \geq u) \leq e \cdot \exp(-\psi_{t_0 s_0}^*(u))$.

A proof of Proposition 15 appears in Section 4.C. In view of Theorem 14, Proposition 15(i) is obtained through an application of the Cramér-Chernoff technique, together with Ville's inequality for reverse submartingales (Theorem 13). It can thus also be seen as an extension of the supermartingale techniques in Howard and Ramdas (2022) to the reversed setting. In Proposition 15(ii), an analogous bound is obtained for the two-sample case, though with the additional factor e in the probability bound. The presence of such a factor greater than 1 is necessary due to the aforementioned counterexample of Cairoli (1970) regarding maximal inequalities for partially ordered martingales, though we do not know if the value e is sharp. The bound itself is obtained via Proposition 13.

Inverting the probability inequalities of Proposition 15 leads to one- and two-sample confidence sequences for $\Phi(P)$ or $\Psi(P, Q)$, though with lengths which are constant with respect to t, s. To obtain confidence sequences scaling at rate-optimal lengths, we employ a stitching construction inspired by those described in Section 4.2.5, together with Proposition 15. Our result will depend on user-specified functions $\ell, g : [0, \infty) \to [1, \infty)$, known as stitching functions, which dictate the shape of the resulting confidence sequences below. We construct these functions to satisfy

$$\sum_{k=1}^{\infty} \frac{1}{\ell(k)} \leq 1, \quad \sum_{j,k=1}^{\infty} \frac{e}{g(k+j)} \leq 1, \tag{4.18}$$

as well as $\ell(j) = \ell(1)$ for all $j \in [0, 1]$, and $g(k) = g(2)$ for all $k \in [0, 2]$. Typical choices include $\ell(k) = (1 \vee k)^{\alpha} \zeta(\alpha)$ and $g(k) = e(2 \vee k)^{\alpha+1}(\zeta(\alpha) - \zeta(\alpha+1))$ (Borwein and Borwein (1987), p. 305), where $\alpha > 1$ and $\zeta(\alpha) = \sum_{k=1}^{\infty}(1/k^{\alpha})$. For ease of exposition, we also insist that the sequences $2^{-u} \log \ell(u)$ and $(2^{-u} + 2^{-v}) \log g(u + v)$ are chosen to be decreasing in each of the indices $u, v \geq 1$, implying that $\ell(u), g(u) = o(\exp(\exp(u)))$. Our main result is below.

Theorem 15. *Let Φ, Ψ denote convex functionals for which the processes (M_{ts}) and (N_t) satisfy the bounds of equation* (4.17)*. For any integer $t \geq 1$, let $\bar{t} = \lceil t/2 \rceil$, and fix $\delta \in (0,1)$.*

(i) (One-Sample) Assume ψ_t^* is invertible for all $t \geq 1$, and that the sequence

$$\gamma_t = (\psi_{\bar{t}}^*)^{-1}\Big(\log \ell(\log_2 t) + \log(2/\delta)\Big), \quad t \geq 1$$

is nonincreasing. Then,

$$\mathbb{P}\Big\{\exists t \geq 1 : \Phi(P_t) \geq \Phi(P) + \gamma_t\Big\} \leq \delta/2.$$

(ii) (Two-Sample) Assume ψ_{ts}^* is invertible for all $t, s \geq 1$, and that the sequence

$$\gamma_{ts} = (\psi_{\bar{t}\bar{s}}^*)^{-1}\Big(\log g(\log_2 t + \log_2 s) + \log(2/\delta)\Big), \quad t, s \geq 1$$

is nonincreasing with respect to the partial order on $\mathbb{N}^2$. Then,

$$\mathbb{P}\Big\{\exists t, s \geq 1 : \Psi(P_t, Q_s) \geq \Psi(P, Q) + \gamma_{ts}\Big\} \leq \delta/2.$$

We begin by noting that for a fixed sample size n, if we denote $\bar{\gamma}_n = (\psi_n^*)^{-1}(\log(2/\delta))$, then the fixed-time Cramér-Chernoff concentration bound corresponding to part (i) is given by (Boucheron, Lugosi, and Massart, 2013),

$$\mathbb{P}\Big\{\Phi(P_n) \geq \Phi(P) + \bar{\gamma}_n\Big\} \leq \delta/2,$$

and an analogous statement can also be made for part (ii) above. Thus, our time-uniform bounds are essentially an iterated logarithm factor worse than the usual fixed-time bounds, but now also apply at arbitrary stopping times.

Before further commenting on the above result, we instantiate it in the special case where N_t and M_{ts} are sub-Gaussian for all $t, s \geq 1$. In Section 4.4, we illustrate how such a condition can be satisfied under tail assumptions on the distributions P and Q themselves.

Corollary 6. *Fix $\delta \in (0,1)$, and recall that $\bar{t} = \lceil t/2 \rceil$.*

(i) (One-Sample) Assume N_t is κ_t^2-sub-Gaussian for some $\kappa_t > 0$, and for all $t \geq 1$. Choose ℓ so that $(\kappa_{\bar{t}}^2 \log \ell(\log_2 t))_{t \geq 1}$ is nonincreasing. Then,

$$\mathbb{P}\left\{\exists t \geq 1 : N_t \geq \mathbb{E}\left(N_{\bar{t}}\right) + \sqrt{2\kappa_{\bar{t}}^2\Big[\log \ell(\log_2 t) + \log(2/\delta)\Big]}\right\} \leq \delta/2,$$

(ii) *(Two-Sample) Assume M_{ts} is σ_{ts}^2-sub-Gaussian for some $\sigma_{ts} > 0$, and for all $s, t \geq 1$. Choose g so that $(\sigma_{\bar{t}\bar{s}}^2 \log g(\log_2 t + \log_2 s))_{t,s\geq 1}$ is nonincreasing. Then,*

$$\mathbb{P}\left\{\exists t, s \geq 1 : M_{ts} \geq \mathbb{E}\left(M_{\bar{t}\bar{s}}\right) + \sqrt{2\sigma_{\bar{t}\bar{s}}^2\Big[\log g(\log_2 t + \log_2 s) + \log(2/\delta)\Big]}\right\} \leq \delta/2.$$

Theorem 15(i) is proved by dividing time $t \geq 1$ into geometrically increasing epochs of the form $[2^j, 2^{j+1}]$, $j \geq 0$, over each of which we construct confidence boundaries at the level $\delta_j/2 = \delta/(2\ell(j+1)) \in (0,1)$ using Proposition 15. Taking a union bound over these boundaries leads to a miscoverage probability of at most $\sum_{j=0}^\infty (\delta_j/2) \leq \delta/2$. The two-sample process (M_{ts}) is handled similarly, by instead forming two sequences of epochs. In Section 4.C, we also state and prove a more general version of Theorem 15 in terms of epoch sizes different than 2, which will be needed in Section 4.4.8.

Remark. Given a convex divergence D, Theorem 15(i) implies that $(1 - \delta/2)$-upper confidence sequences for the processes $N_t^X = D(P_t\|P)$ and $N_s^Y = D(Q_s\|Q)$ are respectively given by

$$\gamma_t^X = (\psi_{X,\bar{t}}^*)^{-1}\Big(\log \ell(\log_2 t) + \log(2/\delta)\Big), \quad \gamma_s^Y = (\psi_{Y,\bar{s}}^*)^{-1}\Big(\log \ell(\log_2 s) + \log(2/\delta)\Big),$$

where $\psi_{X,t}$ is an upper bound on the cumulant generating function of N_t^X, and similarly for $\psi_{Y,s}$. When D satisfies the triangle inequality, one may deduce the following two-sided confidence sequence for $D(P\|Q)$,

$$\begin{aligned}\mathbb{P}\big(\forall t, s \geq 1 :& |D(P_t\|Q_s) - D(P\|Q)| \leq \gamma_t^X + \gamma_s^Y\big) \\ &\geq 1 - \mathbb{P}\big(\exists t \geq 1 : N_t^X > \gamma_t^X\big) - \mathbb{P}\big(\exists s \geq 1 : N_s^Y > \gamma_s^Y\big) \geq 1 - \delta. \end{aligned} \tag{4.19}$$

Equation (4.19) is significant in that it provides a time-uniform bound for a partially ordered reverse submartingale on the basis of two totally ordered reverse submartingales. Doing so bypasses the nearly unavoidable factor e in condition (4.18), but may nevertheless be looser in general due to the application of the triangle inequality. We also remark that equation (4.19) does not require P_t to be independent of Q_s, unlike Theorem 15(ii).

4.3.2 Upper Confidence Sequences via Affine Minorants

The lower confidence sequences derived in Theorem 15 hinged upon the reverse submartingale property of the processes (N_t) and (M_{ts})—an inherently one-sided condition. We show in this section how a different approach can be used to derive upper confidence sequences, motivated both by technical and statistical considerations.

- On the technical side, it would seem natural to repeat the steps of Theorem 15 with respect to the process $(-N_t)$ to obtain an upper confidence sequence. However, $(-N_t)$ is a reverse *super*martingale and thus cannot satisfy infinite-horizon (Ville-type) maximal inequalities. Furthermore, the exponential process $(\exp(-N_t))$ may generally be neither a reverse supermartingale nor a submartingale, thus an analogue of Proposition 15 cannot be derived.

- On the statistical side, the plugin estimators $\Phi(P_t)$ and $\Psi(P_t, Q_s)$ are typically upward biased in estimating $\Phi(P)$ and $\Psi(P, Q)$ respectively. This fact can be deduced from Corollary 7 below, but can already be anticipated from the fact that $\mathbb{E}[\Phi(P_t)]$ and $\mathbb{E}[\Psi(P_t, Q_s)]$ are *nonincreasing* sequences, since $(\Phi(P_t))$ and $(\Psi(P_t, Q_s))$ are reverse submartingales (Theorem 14). This upward bias suggests that confidence sequences for $D(P\|Q)$ of the form $[D(P_t, Q) - \ell_t, D(P_t, Q) + u_t]$ should typically be asymmetric, with the sequence (u_t) potentially decaying at a faster rate than (ℓ_t).

Our approach is summarized as follows. The convexity of Φ guarantees that it can be minorized by an affine functional on $\mathcal{P}(\mathcal{X})$. Notice that an affine functional evaluated at the empirical measure is a sample average, and therefore a reverse martingale. Furthermore, when a convex duality result guarantees that Φ is equal to the supremum over a set of minorizing affine functionals, it can be shown that the difference $\Phi(P_t) - \Phi(P)$ is in fact minorized by a *mean-zero* sample average, for which confidence sequences of length $O(\sqrt{\log\log t/t})$ can be obtained in a standard way under appropriate tail conditions. Doing so leads to a time-uniform lower bound on $\Phi(P_t) - \Phi(P)$, which is easily rephrased as an upper confidence sequence for $\Phi(P)$ scaling at a near-parametric rate.

While this intuition can be made rigorous for a broad collection of convex functionals, we shall avoid doing so in full generality to avoid introducing additional terminology. We shall instead assume that Φ and Ψ take the following form, which is sufficiently general to cover the divergences of primary interest in our development,

$$\Phi(\mu) = \sup_{f \in \mathcal{F}_\Phi} \int f d\mu, \quad \Psi(\mu, \nu) = \sup_{(f,g) \in \mathcal{H}_\Psi} \int f d\mu + \int g d\nu, \tag{4.20}$$

for all $\mu, \nu \in \mathcal{P}(\mathcal{X})$. Here, $\mathcal{F}_\Phi$ denotes a set of Borel-measurable functions on $\mathcal{X}$, $\mathcal{H}_\Psi$ a set of pairs of such functions, and we also write $\mathcal{F}_\Psi = \{f : (f, g) \in \mathcal{H}_\Psi\}$ and $\mathcal{G}_\Psi = \{g : (f, g) \in \mathcal{H}_\Psi\}$. It is clear from equations (4.6), (1.9) and (4.8) that if D is an IPM, φ-divergence, or optimal transport cost, then the functionals $\Psi = D$ and $\Phi = D(\cdot\|Q)$ (for a fixed measure $Q \in \mathcal{P}(\mathcal{X})$) admit the representation (4.20)—for instance, in the case of IPMs generated by a function class $\mathcal{J}$, one may take $\mathcal{F}_\Phi = \{f - \int f dQ : f \in \mathcal{J}\}$ and $\mathcal{H}_\Psi = \{(f, -f) : f \in \mathcal{J}\}$. With this notation in place, the following observation is straightforward.

Proposition 16. If Φ and Ψ admit the representation (4.20), and the suprema therein are achieved for $\mu = P$ and $\nu = Q$, respectively by $f_\Phi \in \mathcal{F}_\Phi$ and by $(f_\Psi, g_\Psi) \in \mathcal{H}_\Psi$, then

1. The process $(N_t)_{t\geq 1}$ is bounded below by

$$R_t = \int f_\Phi d(P_t - P),$$

which is a (mean-zero) reverse martingale with respect to the exchangeable filtration $(\mathcal{E}_t^X)$.

2. The process $(M_{ts})_{t,s\geq 1}$ is bounded below by $(R_t^X + R_s^Y)_{t,s\geq 1}$, where

$$R_t^X = \int f_\Psi d(P_t - P),\ t \geq 1, \quad \text{and} \quad R_s^Y = \int g_\Psi d(Q_s - Q),\ s \geq 1,$$

are (mean-zero) reverse martingales with respect to $(\mathcal{E}_t^X)$ and $(\mathcal{E}_s^Y)$ respectively.

We begin by noting that Proposition 16 implies the aforementioned upward bias of the plugin estimators $\Phi(P_t), \Psi(P_t, Q_s)$, including at arbitrary stopping times. Indeed, by the optional stopping theorem (see for instance Durrett (2019), Theorem 4.8.3.), we can easily infer the following fact that we record formally for reference.

Corollary 7. *Assume the same conditions as Proposition 16, and that the processes $(R_t), (R_t^X), (R_s^Y)$ therein are uniformly integrable. Then, for any stopping times τ and σ with respect to the canonical forward filtrations $(\sigma(X_1, \dots, X_t))_{t=1}^\infty$ and $(\sigma(Y_1, \dots, Y_s))_{s=1}^\infty$ respectively, we have*

$$\mathbb{E}[\Phi(P_\tau)] \geq \Phi(P), \quad \mathbb{E}[\Psi(P_\tau, Q_\sigma)] \geq \Psi(P, Q). \tag{4.21}$$

Furthermore, Proposition 16 can readily be used to form upper confidence sequences for $\Phi(P)$ or $\Psi(P, Q)$ on the basis of the reverse martingales (R_t), or (R_t^X) and (R_s^Y). These processes are simply sample averages, hence they can already be controlled using the existing literature on sequential mean estimation (summarized in Section 4.2.5). Nevertheless, for the purpose of being self-contained, we use reverse martingale techniques to derive such upper confidence sequences under tail conditions on $\mathcal{F}_\Phi$ and $\mathcal{H}_\Psi$, in Proposition 21 of Section 4.C.2. The following special case of this result will be used repeatedly in Section 4.4.

Corollary 8. *Assume the same conditions as Proposition 16. Assume further that for any $h \in \mathcal{F}_\Psi \cup \mathcal{G}_\Psi$, $\mathrm{diam}(h(\mathcal{X})) \leq B < \infty$, and define*

$$\kappa_t = \sqrt{\frac{\log \ell(\log_2 t) + \log(4/\delta)}{t}}, \quad \kappa_{ts} = \kappa_t + \kappa_s. \tag{4.22}$$

Then, for any $\delta \in (0, 1)$, we have $\mathbb{P}(\exists t, s \geq 1 : \Psi(P_t, Q_s) \leq \Psi(P, Q) - B\kappa_{ts}) \leq \delta/2$.

To summarize, under the assumptions and notation of Theorem 15 and Proposition 8, we deduce that a two-sided, two-sample, $(1 - \delta)$-confidence sequence for $\Psi(P, Q)$ is given by

$$C_{ts} = \Big[\Psi(P_t, Q_s) - \kappa_{ts}, \Psi(P_t, Q_s) + \gamma_t^X + \gamma_s^Y\Big], \tag{4.23}$$

and similarly for the functional Φ.

4.3.3 On the Choice of Filtrations and Stopping Times

The majority of confidence sequences derived in past literature, such as those described in Section 4.2.5, employ martingales with respect to the canonical, or "data-generating" filtration. A notable exception is the work of Vovk (2021), which shows that the power of certain sequential tests can be increased by coarsening the canonical filtration. It was similarly fruitful in our work to distinguish the data-generating filtration, with respect to which our processes do not appear to admit any martingale-type property, from a different filtration with respect to which our processes do admit a (reverse) martingale property. To elaborate, let

$$\mathcal{D}_t^X = \sigma(X_1, X_2, \dots, X_t), \quad \mathcal{D}_s^Y = \sigma(Y_1, Y_2, \dots, Y_s), \quad t, s = 1, 2, \dots$$

denote the canonical filtrations associated with each sequence of samples. Our bounds have implicitly assumed that at any pair of times (t, s), the practitioner has access to the information encoded by the data-generating filtration

$$\mathcal{D}_{ts} = \mathcal{D}_t^X \bigvee \mathcal{D}_s^Y, \quad t, s = 1, 2, \dots \tag{4.24}$$

The process (M_{ts}) is naturally adapted to $(\mathcal{D}_{ts})$, but we are not aware of it satisfying a martingale-type property with respect to this filtration in general[2]. It is, however, also adapted to the exchangeable filtration $\mathcal{E}_{ts} = \mathcal{E}_t^X \bigvee \mathcal{E}_s^Y$, but unlike before, (M_{ts}) is also a reverse submartingale with respect to $(\mathcal{E}_{ts})$. Our paper reinforces the somewhat underappreciated view in sequential analysis that filtrations should not be viewed as being "inherent" to the problem, or as tedious formalism for ensuring measurability, but instead viewed as design tools—a nonstandard choice of filtration can yield a powerful design tool.

Validity at Stopping Times. To better understand the underlying role of the filtration $(\mathcal{D}_{ts})$, we shall now prove that the results of Theorem 15 can equivalently be stated as bounds which hold at arbitrary stopping times with respect to $(\mathcal{D}_{ts})$. We focus on the two-sample case in what follows.

In order to define a notion of stopping time which is suitable for our purposes, define the set

$$\overline{\mathbb{N}}^2 = \mathbb{N}^2 \cup \{(t, \infty) : t \geq 1\} \cup \{(\infty, s) : s \geq 1\} \cup \{(\infty, \infty)\},$$

for some symbols $(t, \infty), (\infty, s), (\infty, \infty)$. We endow $\overline{\mathbb{N}}^2$ with the natural partial order, given by that of $\mathbb{N}^2$ (described in Section 4.2.4), together with the following additional relations: $(t, s) \leq (t', \infty)$ whenever $t \leq t'$ and $s \in \mathbb{N}$; $(t, s) \leq (\infty, s')$ whenever $s \leq s'$ and $t \in \mathbb{N}$; $u \leq (\infty, \infty)$ for all $u \in \overline{\mathbb{N}}^2$. A map $\eta : \Omega \to \overline{\mathbb{N}}^2$ is said to be a stopping time with respect to a filtration $(\mathcal{F}_{ts})$ if $\{\eta = (t, s)\} \in \mathcal{F}_{ts}$ for all $(t, s) \in \mathbb{N}^2$.

Intuitively, the event $\{\eta = (t, s)\}$ indicates that the data collection from each of P and Q was terminated at times (t, s), whereas the event $\{\eta = (t, \infty)\}$ indicates that data was collected from P until time t, but indefinitely so from Q. Likewise, the event $\{\eta = (\infty, \infty)\}$ indicates that neither of the two data collections were halted. With these definitions in place, we arrive at the following general equivalence.

Proposition 17. Let $(A_{ts})_{t,s=1}^\infty$ be a sequence of events adapted to a forward filtration $(\mathcal{F}_{ts})_{t,s=1}^\infty$. Define for all $t, s \geq 1$,

$$A_{t\infty} = \limsup_{s\to\infty} A_{ts}, \quad A_{\infty s} = \limsup_{t\to\infty} A_{ts}, \quad A_{\infty\infty} = \left(\limsup_{t\to\infty} A_{t\infty}\right) \cup \left(\limsup_{s\to\infty} A_{\infty s}\right). \tag{4.25}$$

Then, for all $\delta \in (0, 1)$, the following statements are equivalent.

(i) $\mathbb{P}\left(\bigcup_{t,s=1}^\infty A_{ts}\right) \leq \delta$.

[2] Nevertheless, under some conditions, it can be deduced from Theorem 14 that there exists a bivariate canonical filtration $(\mathcal{F}_{ts})$ and a process $(\widetilde{M}_{ts})$, which has the same distribution as (M_{ts}), such that $(\widetilde{M}_{ts})$ is a reverse submartingale with respect to $(\mathcal{F}_{ts})$ rather than $(\mathcal{E}_{ts})$—see Theorem B of Rzeszut and Trojan (2020).

(ii) For any stopping time (τ, σ) with respect to $(\mathcal{F}_{ts})$, we have $\mathbb{P}(A_{\tau\sigma}) \leq \delta$.

(iii) For any random time (T, S), not necessarily a stopping time, we have $\mathbb{P}(A_{TS}) \leq \delta$.

The proof of Proposition 17 is given in Section 4.C.3. Analogues of Proposition 17 for one-sample processes have previously been given by Howard et al. (2021), Ramdas et al. (2020), and Zhao et al. (2016), so our result is an extension of theirs to partially ordered processes. In our setting, recall that M_{ts} is $(\mathcal{D}_{ts} \bigwedge \mathcal{E}_{ts})$-measurable. While Proposition 17 could be reformulated in reverse time, so that $(\mathcal{F}_{ts})$ can be taken to be the modeling filtration $(\mathcal{E}_{ts})$, it is most interpretable to take it to be the data-generating filtration $(\mathcal{D}_{ts})$. Doing so, under the assumptions of Theorem 15, leads for instance to the bound

$$\mathbb{P}\big\{\Psi(P,Q) \in C_{\tau\sigma}\big\} \geq 1-\delta \quad \text{for all stopping times } \eta = (\tau, \sigma) \text{ with respect to } (\mathcal{D}_{ts}), \tag{4.26}$$

where C_{ts} denotes the two-sided interval (4.23), understood with conventions for infinities which can be deduced from equation (4.25).

Alternate Data-Generating Filtrations. Though we presumed the data-generating filtration (4.24) throughout our development, slightly tighter confidence sequences can be obtained if the user has access to additional information. For instance, our confidence sequences hold uniformly over arbitrary pairs of time (t, s), but such flexibility is unnecessary if the practitioner knows the order in which sample points from P and Q arrive. We illustrate two such examples below, focusing on lower confidence sequences:

(i) **Paired Samples.** When the observations X_t and Y_t are presumed to arrive at the same time, in pairs (X_t, Y_t), the data-generating filtration may be replaced by

$$\mathcal{D}_t = \sigma(X_t, Y_t), \quad t = 1, 2, \dots$$

In this case, following along similar lines as before, the following two-sample bound may be established, and is tighter than that of Theorem 15(ii),

$$\mathbb{P}\left\{\exists t \geq 1 : \Psi(P_t, Q_t) \geq \Psi(P,Q) + (\psi_{tt}^*)^{-1}\big(\log \ell(\log_2 t) + \log(1/\delta)\big)\right\} \leq \delta. \tag{4.27}$$

Unlike Theorem 15, we note that the bound (4.27) can be taken to hold without assuming that (X_t) and (Y_s) are independent of each other.

(ii) **Samples Ordered by External Randomization.** As a generalization of the previous point, assume the observations X_t and Y_s arrive in a possibly random order which is independent of the data. Specifically, let $(\iota_n)_{n\geq 1}$ denote a sequence of random variables taking values in $\{0, 1\}$, which are independent of (X_t) and (Y_s), but possibly dependent on an external source of randomness U, say distributed uniformly on $[0, 1]$. Let $t(n) = \sum_{i=1}^n \iota_n$, and $s(n) = n - t_n$ so that at any time $n \geq 1$, the practitioner observes $\iota_n X_{t(n)} + (1-\iota_n) Y_{s(n)}$. In this case, one has access to the filtration $\mathcal{I}_n = \sigma(U, \iota_1, \iota_2, \dots, \iota_n)$, $n \geq 1$, which determines the order in which the sample points X_t, Y_s arrive, as well as to the data-generating filtration

$$\overline{\mathcal{D}}_n = \overline{\mathcal{D}}_{t(n)}^X \bigvee \overline{\mathcal{D}}_{s(n)}^Y, \quad n \geq 1, \tag{4.28}$$

where $\overline{\mathcal{D}}_{t(n)}^X$ and $\overline{\mathcal{D}}_{s(n)}^Y$ are defined similarly as follows:

$$\overline{\mathcal{D}}_{t(n)}^X = \{A \in \mathcal{F} : A \cap \{t(n) = t\} \in \overline{\mathcal{D}}_t^X, \ \forall t \geq 1\}, \quad \text{where} \quad \overline{\mathcal{D}}_t^X = \sigma(U, X_1, \dots, X_t). \tag{4.29}$$

Note that we could have assumed that the sequence (ι_n) is fully deterministic, in exchange for simpler notation. However, there are many situations, like clinical trials, in which we may wish to use external randomization (encoded by U) to determine how to obtain the next data point; for example, Efron (1971) shows how to adaptively randomize participants while encouraging balance between $t(n)$ and $s(n)$. Under this setting, it can be shown that $(\Psi(P_{t(n)}, Q_{s(n)}))_{n\geq 1}$ is a reverse submartingale, and assuming for simplicity that $\psi_n = \psi_{n0} = \psi_{0n}$, one may obtain the confidence sequence

$$\mathbb{P}\big\{\exists n \geq 1 : \Psi(P_{t(n)}, Q_{s(n)}) \geq \Psi(P, Q) + (\psi_{\bar{n}}^*)^{-1}\big(\log \ell(\log_2 n) + \log(1/\delta)\big)\big\} \leq \delta.$$

In contrast to the above two settings, our confidence sequence C_{ts} satisfies the guarantee (4.3), which is uniform over pairs $(t, s) \in \mathbb{N}^2$, and therefore yields valid coverage even if the orderings $t(n)$ and $s(n)$ depend arbitrarily on the samples observed from P and Q.

4.4 Explicit Corollaries for Common Divergences

We now specialize the confidence sequence C_{ts} to several examples of divergences including IPMs (Sections 4.4.1, 4.4.2), optimal transport costs (Section 4.4.3), φ-divergences (Section 4.4.4), and divergences smoothed by convolution (Section 4.4.5). Moving beyond divergences, we also derive time-uniform generalization error bounds for binary classification problems (Section 4.4.6), and confidence sequences for multivariate means (Section 4.4.7). These special cases will illustrate how our framework can be used to port existing fixed-time concentration inequalities to time-uniform ones, typically at the expense of iterated logarithmic factors. In these cases, any improvements to existing fixed-time concentration inequalities would typically carry over to our time-uniform setting. Though our focus is on non-asymptotic bounds, in Section 4.4.8, we also show how Theorem 15 can be used to derive a one-sided analogue of the classical law of the iterated logarithm, for several divergences between an empirical and true underlying measure. We defer all proofs to Section 4.D.

4.4.1 Kolmogorov-Smirnov Distance

Theorem 15 leads to a sequential analogue of the classical Dvoretzky-Kiefer-Wolfowitz (DKW) inequality (Dvoretzky, Kiefer, and Wolfowitz, 1956; Massart, 1990), based on distinct techniques than those of Howard and Ramdas (2022) and Maillard (2021). Let P be any distribution over $\mathbb{R}$ with cumulative distribution function (CDF) F. Let $F_t(x) = (1/t)\sum_{i=1}^n I(X_i \leq x)$ denote the empirical CDF of F.

Corollary 9. *For any $\delta \in (0, 1)$,*

$$\mathbb{P}\left(\exists t \geq 1 : \|F_t - F\|_\infty \geq \sqrt{\frac{\pi}{t}} + 2\sqrt{\frac{2}{t}\Big[\log \ell(\log_2 t) + \log(1/\delta)\Big]}\right) \leq \delta.$$

Notice that Corollary 9 involves the term $\log(1/\delta)$, as opposed to the term $\log(2/\delta)$ which appears in the classical DKW inequality. This is due to the one-sidedness of the bounds in Theorem 15. The price to pay is the additional additive term $\sqrt{\pi/t}$, which is an upper bound on the expectation of $\|F_{\lceil t/2\rceil} - F\|_\infty$.

Corollary 9 and Proposition 21 give rise to a sequential analogue of the celebrated Kolmogorov-Smirnov two-sample test. In what follows, Q denotes a second distribution over $\mathbb{R}$ with CDF G, and empirical CDF $G_s(y) = (1/s)\sum_{i=1}^s I(Y_i \leq y)$. Recall also the sequence (κ_{ts}) defined in equation (4.22).

Corollary 10. *Given $\delta \in (0,1)$, set*

$$\gamma_{ts} = \sqrt{\pi/t} + \sqrt{\pi/s} + 2\sqrt{\frac{2ts}{t+s}\big[\log g(\log_2 t + \log_2 s) + \log(2/\delta)\big]}.$$

Then,

$$\mathbb{P}\big(\forall t, s \geq 1 : -\gamma_{ts} \leq \|F - G\|_\infty - \|F_t - G_s\|_\infty \leq \kappa_{ts}\big) \geq 1 - \delta.$$

In particular, the sequential Kolmogorov-Smirnov test which rejects the null hypothesis $H_0 : P = Q$ when $\|F_t - G_s\|_\infty > \gamma_{ts}$ has type-I error controlled at $\delta/2$.

We now turn our attention to another popular IPM that is based on reproducing kernels.

4.4.2 Maximum Mean Discrepancy, V-Statistics, and U-Statistics

The kernel Maximum Mean Discrepancy (MMD) is an IPM measuring the distance between embeddings of distributions in a reproducing kernel Hilbert space (RKHS). We provide a brief definitions in what follows, and refer the reader to Schölkopf and Smola (2018) for further details. Let $K : \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}_+$ be a Mercer kernel, that is a symmetric and continuous function such that for any finite set of points $x_1, \dots, x_n \in \mathbb{R}^d$, the matrix $(K(x_i, x_j))_{i,j=1}^n$ is positive semidefinite. The RKHS $\mathcal{H}$ corresponding to K is the closure of the set

$$\mathcal{H}_0 = \left\{\sum_{i=1}^k \alpha_i K(\cdot, x_i) : \alpha_i \in \mathbb{R},\ x_i \in \mathbb{R}^d, k \geq 1\right\},$$

endowed with the inner product and norm

$$\langle f, g\rangle_{\mathcal{H}} = \sum_{i=1}^k \sum_{j=1}^{k'} \alpha_i \beta_j K(x_i, y_j), \quad \|f\|_{\mathcal{H}} = \sqrt{\langle f, f\rangle_{\mathcal{H}}},$$

where $f = \sum_{i=1}^k \alpha_i K(\cdot, x_i)$ and $g = \sum_{j=1}^{k'} \beta_j K(\cdot, y_j)$ denote the expansions of any two functions $f, g \in \mathcal{H}_0$. The MMD is defined as the IPM over the unit ball $\mathcal{J} = \{f \in \mathcal{H} : \|f\|_{\mathcal{H}} \leq 1\}$ of $\mathcal{H}$. The plugin estimator of the MMD admits the following representation, which makes its computation straightforward

$$\mathrm{MMD}(P_t, Q_s) = \sqrt{\frac{1}{t^2}\sum_{i,j=1}^t K(X_i, X_j) + \frac{1}{s^2}\sum_{i,j=1}^s K(Y_i, Y_j) - \frac{2}{st}\sum_{i=1}^t\sum_{j=1}^s K(X_i, Y_j)}. \tag{4.30}$$

In particular, if $s = t$ and the data are available in pairs $Z_t = (X_t, Y_t)$ for all $t \geq 1$, as in Section 4.3.3(1), the above expression may be rewritten as a square root of a second-order V-Statistic,

$$\mathrm{MMD}(P_t, Q_t) = \sqrt{\frac{1}{t^2} \sum_{i,j=1}^{t} J(Z_i, Z_j)}, \tag{4.31}$$

where $J((x,y),(x',y')) = K(x,x') + K(y,y') - 2K(x,y')$, for any $x, x', y, y' \in \mathbb{R}^d$. Assuming the kernel K is bounded, we derive a sequential concentration bound for this statistic as follows.

Corollary 11. *Let $P, Q \in \mathcal{P}(\mathbb{R}^d)$. Assume that $\sup\{K(x,y) : x, y \in \mathbb{R}^d\} \leq B < \infty$. For any $\delta \in (0,1)$, define*

$$\gamma_{ts} = 2\sqrt{2B}(t^{-\frac{1}{2}} + s^{-\frac{1}{2}}) + 4\sqrt{\frac{B(t+s)}{ts}\Big[\log g(\log_2 t + \log_2 s) + \log(2/\delta)\Big]},$$

and let (κ_{ts}) be the sequence defined in equation (4.22). Then,

$$\mathbb{P}\Big(\forall t, s \geq 1 : -\gamma_{ts} \leq \mathrm{MMD}(P,Q) - \mathrm{MMD}(P_t, Q_s) \leq 2\sqrt{B}\kappa_{ts}\Big) \geq 1 - \delta.$$

Assuming that the stitching function g is bounded above by a polynomial, Corollary 11 provides a confidence sequence for $\mathrm{MMD}(P,Q)$ scaling at the rate $O(\sqrt{\log\log(t \vee s)/(t \vee s)})$. Up to the iterated logarithmic factor, we recover the fixed-time rate obtained by Gretton et al. (2012) (Theorem 7), which was shown to be minimax optimal by Tolstikhin, Sriperumbudur, and Schölkopf (2016). In the setting of equal sample sizes $t = s$ (cf. Section 4.3.3(1)), it is well-known that the V-Statistic $\mathrm{MMD}^2(P_t, Q_t)$ has first-order degeneracy, so that $\mathrm{MMD}^2(P_t, Q_t) = O_p(1/t)$ when $P = Q$ (Lee, 1990). On the other hand, the bound $|\mathrm{MMD}^2(P_t, Q_t) - \mathrm{MMD}^2(P,Q)| = O_p(1/\sqrt{t})$ is tight when P, Q are fixed and $P \neq Q$. On the squared scale, the bound of Corollary 11 adapts to these distinct rates of convergence, since it implies that with probability at least $1 - \delta$,

$$\forall t \geq 1, \quad |\mathrm{MMD}^2(P_t, Q_t) - \mathrm{MMD}^2(P,Q)| = O\left(\mathrm{MMD}(P,Q)\sqrt{\frac{\log\log t}{t}} + \frac{\log\log t}{t}\right). \tag{4.32}$$

Above, the right-hand side decays at the rate $O(\sqrt{(\log\log t)/t})$ in general, but improves to $O\left((\log\log t)/t\right)$ when $P = Q$. Similar considerations are discussed by Gretton et al. (2012).

While the plugin estimator $\mathrm{MMD}(P_t, Q_s)$ is typically upwards biased, the *squared* MMD also admits a widely-used unbiased estimator (Gretton et al., 2012) obtained by replacing the V-Statistics in equations (4.30) and (4.31) by U-Statistics. We derive confidence sequences for $\mathrm{MMD}^2(P,Q)$ based on this estimator in Section 4.D.2. The bounds therein do not adapt to the distinct rates of convergence described above, however, therefore we recommend those of Corollary 11 in practice.

We conclude this section with a more general discussion of sequential inference based on U- and V-Statistics, for expectation functionals of the form

$$\Phi : \mathcal{P}(\mathbb{R}^d) \to \mathbb{R}, \quad \Phi(P) = \iint h(x,y)dP(x)dP(y),$$

where $h : \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}$ is a symmetric function. Following Lee (1990); Serfling (2009), the U- and V-Statistics corresponding to Φ are respectively defined by

$$U_t = \mathbb{E}[h(X_1, X_2)|\mathcal{E}_t^X] = \frac{2}{t(t-1)} \sum_{1\leq i<j\leq t} h(X_i, X_j), \quad V_t = \Phi(P_t) = \frac{1}{t^2} \sum_{i,j=1}^{t} h(X_i, X_j).$$

The above representation immediately implies that (U_t) is a reverse martingale whenever it is integrable—a fact which can equivalently be derived from the leave-one-out property (4.16), which holds for (U_t) with equality. Notice that this property holds irrespective of the kernel h, so long as it is symmetric. On the other hand, the following can be said about (V_t).

Proposition 18. Assume that $h : \mathcal{X} \times \mathcal{X} \to \mathbb{R}_+$ is a continuous, symmetric, positive definite kernel over a compact set $\mathcal{X} \subseteq \mathbb{R}^d$. Then $\sqrt{\Phi(\cdot)}$ is a convex functional, thus $(\sqrt{V_t})$ and (V_t) are reverse submartingales with respect to $(\mathcal{E}_t^X)$.

The proof is in Section 4.D.2. We do not generally expect the above result to hold for any symmetric kernel h, because the functional Φ is akin to a quadratic form, which may be nonconvex if its kernel is not positive semidefinite. Under the above results, a straightforward generalization of Theorem 15 can be used to derive two-sided confidence sequences for $\Phi(P)$ centered at U_t for all symmetric h, or lower confidence sequences for $\sqrt{\Phi(P)}$ and $\Phi(P)$ based on (V_t) for all positive definite h (which may be coupled with upper confidence sequences similarly as in Section 4.3.2). While these considerations make (U_t)-based confidence sequences seem attractive, we recall that (V_t)-based confidence sequences sometimes have the advantage of providing rate-optimal inference even when Φ is degenerate.

4.4.3 Optimal Transport Costs

Let $c : \mathcal{X} \times \mathcal{X} \to \mathbb{R}_+$ be a nonnegative cost function, and assume for simplicity that c is bounded above over $\mathcal{X}$ by $\Delta := \sup_{x,y\in\mathcal{X}} c(x, y) < \infty$. We derive confidence sequences for the optimal transport cost $\mathcal{T}_c(P, Q)$, which will depend on an upper bound $\alpha_{c,ts}$ for the bias of the empirical plugin estimator,

$$\mathbb{E}\big[\mathcal{T}_c(P_t, Q_s)\big] - \mathcal{T}_c(P, Q) \leq \alpha_{c,ts}. \tag{4.33}$$

For instance, one may take $\alpha_{c,ts}$ to scale as $(t\wedge s)^{-1/2}$ when $\mathcal{X}$ is one-dimensional (cf. Chapter 2), finite (Sommerfeld and Munk, 2018; Forrow et al., 2018) or more generally, of intrinsic dimension less than or equal to three (Hundrieser, Staudt, and Munk, 2022). In these cases, the bias term $\alpha_{c,ts}$ will be negligible in the confidence sequence we construct below. More generally, our results in Chapter 3 imply that one may take $\alpha_{c,ts} \lesssim (t \wedge s)^{-\alpha/d}$ for $d \geq 5$ when c is an α-Hölder smooth cost for some $\alpha \in [1, 2]$. In this case, the bias will be of leading order, and our confidence sequence below becomes primarily of theoretical interest.

Corollary 12. *Let $c : \mathcal{X} \times \mathcal{X} \to \mathbb{R}_+$ be a lower semi-continuous cost function bounded above by Δ, assume that equation* (4.33) *is satisfied, and recall that $\bar{t} = \lceil t/2 \rceil$. Furthermore, let (κ_{ts}) be* *the sequence defined in equation* (4.22). *Then, for any* $\delta \in (0,1)$,

$$\mathbb{P}\Bigg(\forall t, s \geq 1 : -\alpha_{c,\bar{t}\bar{s}} - 2\Delta\sqrt{\frac{2(t+s)}{ts}\Big[\log g(\log_2 t + \log_2 s) + \log(2/\delta)\Big]} \\ \leq \mathcal{T}_c(P,Q) - \mathcal{T}_c(P_t,Q_s) \leq \Delta\kappa_{ts}\Bigg) \geq 1-\delta.$$

Corollary 12 exhibits a confidence sequence for $\mathcal{T}_c(P,Q)$ whose length is typically dominated by the expected deviation bound $\alpha_{c,\bar{t}\bar{s}}$. The potentially severe dependence on dimensionality in these rates can limit the applicability of Corollary 12 in high-dimensional problems. Section 4.4.5 derives confidence sequences for smoothed 1-Wasserstein distances, which admit significantly improved dimension dependence.

4.4.4 φ-Divergences over Finite Sets

Let P be a probability distribution supported on a set $\mathcal{X} = \{a_1, \dots, a_k\}$ of finite cardinality $k \geq 2$. Set $p_j = P(\{a_j\})$ for all $j = 1, \dots, k$. In this setting, we write the empirical measure as

$$P_t = \frac{1}{t}\sum_{i=1}^t \delta_{X_i} = \frac{1}{t}\sum_{j=1}^k C_j \delta_{a_j}, \quad \text{where} \quad C_j = \sum_{i=1}^t I(X_i = a_j), \quad j = 1, \dots, k.$$

The vector $(C_1, \dots, C_k)$ can be viewed as a random sample from a multinomial experiment with t trials and k categories with probabilities $(p_1, \dots, p_k)$. Concentration inequalities for φ-divergences D_φ between the empirical measure and the finitely-supported true distribution P have received significant attention in the offline setting. Here, we show how such results can be used together with Theorem 15 to obtain time-uniform bounds on $D_\varphi(P_t\|P)$. We focus on the Kullback-Leibler divergence and Total Variation distance in what follows.

Kullback-Leibler Divergence. Tight upper bounds on the moment generating function of the scaled Kullback-Leibler divergence $t\mathrm{KL}(P_t\|P)$ have recently been derived by Guo and Richardson (2020) (see also Agrawal (2020)), who prove

$$\mathbb{E}[\exp(\lambda t \mathrm{KL}(P_t\|P))] \leq G_{k,t}(\lambda) := \sum_{x_1,\dots,x_k} \binom{t}{x_1, \dots, x_k} \prod_{i=1}^k [\lambda x_i/t + (1-\lambda)p_i]^{x_i}, \tag{4.34}$$

for all $\lambda \in [0,1]$. Guo and Richardson (2020) show that this upper bound is nearly tight, in the sense that it nearly matches the scaling in k and t of the moment generating function of the limiting distribution of $t\mathrm{KL}(P_t\|P)$. Furthermore, the value of $G_{k,t}(\lambda)$ does not depend on P (Guo and Richardson (2020), Proposition 1). Nevertheless, $G_{k,t}$ cannot easily be used in Theorem 15, since the optimization problem $\sup_{\lambda\in[0,1]}\big\{\lambda u - G_{k,t}(\lambda)\big\}$, for any $u > 0$, is non-convex. Guo and Richardson (2020) instead derive several closed-form sequences (λ_t) which approximately solve this maximization problem. We derive a sequential analogue of their bounds in terms of a generic choice of such a sequence. The following result is obtained by repeating a similar stitching argument as that of the proof of Theorem 15.

Proposition 19. Let $\delta \in (0,1)$, and let P be a distribution supported on a finite set of size $k \geq 2$. Let $(\lambda_t)_{t=1}^\infty \subseteq [0,1/2]$ be a sequence of real numbers such that

$$\gamma_t = \frac{2}{\lambda_t t} \log\left(\frac{G_{k,\lfloor t/2 \rfloor}(2\lambda_t)\ell(\log_2 t)}{\delta}\right), \quad t \geq 1,$$

is a nonincreasing sequence in t. Then, $\mathbb{P}\left\{\exists t \geq 1 : \mathrm{KL}(P_t \| P) \geq \gamma_t\right\} \leq \delta$.

It can be seen that the fitted probability vector $(C_1/t, \dots, C_k/t)$ is precisely the maximum likelihood estimator of $(p_1, \dots, p_k)$, and that the scaled Kullback-Leibler divergence $t\mathrm{KL}(P_t \| P)$ is a multiple of the log-likelihood ratio of $(p_1, \dots, p_k)$. Proposition 19 therefore leads to a confidence sequence for the probability vector p on the basis of the classical likelihood ratio statistic.

Total Variation Distance. We now similarly derive time-uniform bounds for the discrete Total Variation distance $\|P_t - P\|_{\mathrm{TV}} := \frac{1}{2}\sum_{j=1}^k |(C_j/t) - p_j|$. The following Corollary follows from Theorem 15 using elementary tail bounds for the Total Variation distance (see for instance Berend and Kontorovich (2013)), together with an expectation bound due to Kamath et al. (2015).

Corollary 13. *For all* $\delta \in (0,1)$, *we have*

$$\mathbb{P}\left\{\exists t \geq 1 : \|P_t - P\|_{\mathrm{TV}} \geq \sqrt{\frac{k}{\pi t}} + 2\left(\frac{2k}{t}\right)^{\frac{3}{4}} + 2\sqrt{\frac{2}{t}\Big[\log \ell(\log_2 t) + \log(1/\delta)\Big]}\right\} \leq \delta.$$

Up to a polylogarithmic factor, the bound of Corollary 13 scales at the parametric rate of convergence when the alphabet size k is fixed. For general distributions with uncountable support, such rates are not achievable under the Total Variation distance due to the lack of absolute continuity of P_t with respect to P. The following subsection studies a notable exception, in which parametric rates are retained when the measures are smoothed by convolution with a kernel admitting fixed bandwidth.

4.4.5 Smoothed Divergences and Differential Entropy

Let $K : \mathbb{R}^d \to \mathbb{R}_+$ denote a smoothing kernel, that is, a nonnegative and continuous function satisfying $\int_{\mathbb{R}^d} K(x)dx = 1$. Given a bandwidth $\sigma > 0$, let $\mathcal{K}_\sigma$ be the probability measure admitting density $K_\sigma(x) = (1/\sigma^d)K(x/\sigma)$ with respect to the Lebesgue measure. Let D denote a convex divergence, and define its smoothed counterpart by

$$D^\sigma : (P,Q) \longmapsto D(P \star \mathcal{K}_\sigma \| Q \star \mathcal{K}_\sigma).$$

It can be directly verified that D^σ is itself a convex divergence, due to the linearity of the convolution operator. Theorem 15 can therefore be used to derive a confidence sequence for $D^\sigma(P\|Q)$ based on the plugin estimator $D^\sigma(P_t\|Q_s)$. We emphasize that this estimator is sensible even if the original divergence D requires absolute continuity of the distributions being compared, as is the case for φ-divergences. In such cases, D^σ forms a proxy of D

which can be estimated using the empirical plugin estimator. We refer to Goldfeld et al. (2020) for upper bounds on $\mathbb{E}D^\sigma(P_t, P)$ under a wide range of divergences D. Smoothing by Gaussian convolution has also recently been studied as a means of regularizing optimal transport problems, and thereby reducing the curse of dimensionality in estimating Wasserstein distances. For instance, Goldfeld and Greenewald (2020), Goldfeld, Greenewald, and Kato (2020) show that that the empirical measure converges to P at the parametric rate under W_1^σ, in expectation, contrasting the unavoidable $t^{-1/d}$ rate for this problem under W_1 itself (Singh and Póczos, 2019).

Motivated by these two applications, we show in what follows how Theorem 15 can be used to derive confidence sequences for the smoothed Total Variation distance, and for the smoothed 1-Wasserstein distance. The results which follow are obtained by first deriving upper bounds on the moment generating functions of $D^\sigma(P_t\|P)$, and second, invoking the upper bounds of Goldfeld et al. (2020) on the expectation of this quantity. In order to appeal to their results, we assume in what follows that $K(x) = e^{-\|x\|_2^2/2}/\sqrt{2\pi}$ is taken to be the standard Gaussian kernel.

We first recall a tail assumption which will be used in the sequel. Given a metric d on $\mathcal{X}$, we say a measure $P \in \mathcal{P}_1(\mathcal{X})$ satisfies the $T_1(\tau^2)$ inequality with respect to d, for some $\tau > 0$, if

$$\mathcal{T}_d(\mu, P) \le \sqrt{2\tau^2 \mathrm{KL}(\mu\|P)}, \quad \text{for all } \mu \in \mathcal{P}_1(\mathcal{X}).$$

Such inequalities are at the heart of the transportation method for deriving fixed-time concentration inequalities—we refer to Gozlan and Léonard (2010) for a survey. For our purposes, transportation inequalities are known to provide a natural tail assumption on P in order to guarantee sub-Gaussian concentration of empirical Wasserstein distances: Niles-Weed and Rigollet (2022) (Theorem 6) prove that P satisfies $T_1(\tau^2)$ if and only if $W_1(P_t, P)$ is (τ^2/t)-sub-Gaussian. By extending their result to smoothed Wasserstein distances, we arrive at the following.

Proposition 20 (Smoothed Divergences). Let $\delta \in (0,1)$, $\sigma > 0$, and let $P \in \mathcal{P}(\mathbb{R}^d)$.

(i) (Total Variation Distance) Assume P is τ^2-sub-Gaussian for some $\tau > 0$. Then,

$$\mathbb{P}\left(\exists t \ge 1 : \|P_t - P\|_{\mathrm{TV}}^\sigma \ge \frac{c_d}{\sqrt{t}} + 4\sqrt{\frac{2}{t}\Big[\log \ell(\log_2 t) + \log(1/\delta)\Big]}\right) \le \delta,$$

where $c_d = \sqrt{2}\left(\frac{1}{\sqrt{2}} + \frac{\tau}{\sigma}\right)^{\frac{d}{2}} e^{\frac{3d}{16}}$.

(ii) (W_1 Distance) Assume P satisfies the $T_1(\tau^2)$ inequality with respect to $\|\cdot\|_2$, for some $\tau > 0$. Then, $W_1^\sigma(P_t, P)$ is (τ^2/t)-sub-Gaussian, and for all $\delta \in (0,1)$,

$$\mathbb{P}\left(\exists t \ge 1 : W_1^\sigma(P_t, P) \ge \frac{C_d}{\sqrt{t}} + 2\sqrt{\frac{\tau^2}{t}\Big[\log \ell(\log_2 t) + \log(1/\delta)\Big]}\right) \le \delta, \tag{4.35}$$

where $C_d = 2\sqrt{d\sigma^2}\left(\frac{1}{\sqrt{2}} + \frac{\tau}{\sigma}\right) c_d$.

Extensions of Proposition 20 to the two-sample setting are straightforward and omitted for brevity. Proposition 20(i) yields a confidence sequence for the smoothed Total Variation under a mere moment condition. Such a result could not have been obtained by our framework in the absence of smoothing, except in the special case of Section 4.4.4 where P was assumed to have countable support. Proposition 20(ii) contrasts our earlier Corollary 12, which implied a confidence sequence for the 1-Wasserstein distance scaling at the rate $O(t^{-1/d})$ for $d \geq 3$. Smoothing removes the dimension dependence from the rate itself when σ is held fixed, although the constant C_d continues to grow exponentially in d. It can be deduced from Theorem 1 of Goldfeld et al. (2020), combined with equation (4.45), below that exponential dimension dependence in this constant is necessary, although the optimal constant is not known. Any sharpening of these constants in future work could directly be used to update the time-uniform bound in Proposition 20.

Inspired by Weed (2018), we briefly close this subsection by illustrating how Proposition 20 can further be used to obtain sequential bounds for the smoothed differential entropy of P, $h(P \star \mathcal{K}_\sigma) = -\int \log(P \star K_\sigma) dP \star \mathcal{K}_\sigma$, using the fact that it is Lipschitz with respect to the W_1 metric (Polyanskiy and Wu, 2016).

Corollary 14. *Let $\delta \in (0,1)$, $\sigma > 0$, and $P \in \mathcal{P}([-1,1]^d)$. Then,*

$$\mathbb{P}\Bigg(\forall t \geq 1 : |h(P_t \star \mathcal{K}_\sigma) - h(P \star \mathcal{K}_\sigma)| \leq \frac{3}{\sigma^2}\sqrt{\frac{d}{t}\Big[\log \ell(\log_2 t) + \log(4/\delta)\Big]} + \frac{\sqrt{d} C_d}{\sqrt{t\sigma^2}}\Bigg) \geq 1 - \delta.$$

Notice that the mapping $P \mapsto -h(P \star \mathcal{K}_\sigma)$ is convex, therefore a confidence sequence for the smoothed differential entropy could also have been obtained by directly appealing to Theorem 15, assuming a bound on the cumulant generating function of $h(P_t \star \mathcal{K}_\sigma)$ were available. Beyond this approach and the bound of Corollary 14, we are not aware of other existing sequential concentration inequalities for this problem.

4.4.6 Sequential Generalization Error Bounds for Binary Classification

Theorem 15 can be used to derive generalization error bounds for classification or regression problems that are valid at stopping times. We illustrate the special case of binary classification. Let $\mathcal{X}$ be a topological space, P be a Borel probability distribution over $\mathcal{X} \times \{-1,1\}$, and $(X_t, Y_t)_{t=1}^\infty$ a sequence of i.i.d. observations from P. Let $\mathcal{F}$ be a collection of Borel-measurable functions from $\mathcal{X}$ to $\{-1,1\}$, and define the population and empirical classification risks by

$$R(f) = \mathbb{P}(f(X) \neq Y), \quad \text{and} \quad R_t(f) = \frac{1}{t}\sum_{i=1}^t I(f(X_i) \neq Y_i), \quad f \in \mathcal{F}.$$

High-probability bounds on the supremum of the empirical process $\sup_{f\in\mathcal{F}} |R(f) - R_t(f)|$ for fixed times $t \geq 1$ are well-studied, and can lead to conservative confidence intervals for the generalization error $R(\widehat{f}_t)$ of any data-dependent classifier $\widehat{f}_t \in \mathcal{F}$, such as an empirical risk minimizer, or an approximate one obtained by stochastic optimization. Such bounds necessarily depend on the complexity of $\mathcal{F}$, as measured for instance by population or empirical Rademacher complexities, respectively defined by

$$\mathcal{R}_t(\mathcal{F}) = \mathbb{E}_{\boldsymbol{\varepsilon}_t,\mathbf{X}_t}\left[\sup_{f\in\mathcal{F}} \frac{1}{t}\left|\sum_{i=1}^t \epsilon_i f(X_i)\right|\right], \quad \widehat{\mathcal{R}}_t(\mathcal{F}) = \mathbb{E}_{\boldsymbol{\varepsilon}_t}\left[\sup_{f\in\mathcal{F}} \frac{1}{t}\left|\sum_{i=1}^t \epsilon_i f(X_i)\right|\right].$$

Here, we denote $\mathbf{X}_t = (X_1, \dots, X_t)$ and $\boldsymbol{\varepsilon}_t = (\epsilon_1, \dots, \epsilon_t)$, where $(\epsilon_t)_{t=1}^\infty$ denotes a sequence of i.i.d. Rademacher random variables (taking values ± 1 with equal probability). We obtain the following bound which is uniform both over hypotheses $f \in \mathcal{F}$ and over time $t \geq 1$.

Corollary 15. *Let $\delta \in (0,1)$, and recall that $\bar{t} = \lfloor t/2 \rfloor$ for all $t \geq 1$.*

1. *The population Rademacher complexity provides a time-uniform generalization error bound:*

$$\mathbb{P}\left(\forall t \geq 1 : \sup_{f\in\mathcal{F}} |R_t(f) - R(f)| \leq \mathcal{R}_{\bar{t}}(\mathcal{F}) + 2\sqrt{\frac{2}{t}\Big[\log \ell(\log_2 t) + \log(1/\delta)\Big]}\right) \geq 1-\delta.$$

2. *The empirical Rademacher complexity $(\widehat{\mathcal{R}}_t(\mathcal{F}))$ is a reverse submartingale with respect to $(\mathcal{E}_t^X)$, and we have:*

$$\mathbb{P}\Big(\forall t \geq 1 : \mathcal{R}_{\bar{t}}(\mathcal{F}) \geq \widehat{\mathcal{R}}_t(\mathcal{F}) - 4\sqrt{\frac{2}{t}\Big[\log \ell(\log_2 t) + \log(1/\delta)\Big]}\Big) \geq 1-\delta.$$

In particular, if τ is an arbitrary stopping time and $\widehat{f}_t$ is an arbitrary data-dependent classifier, then $\mathbb{P}\big(|R_\tau(\widehat{f}_\tau) - R(\widehat{f}_\tau)| \leq \mathcal{R}_{\bar{\tau}}(\mathcal{F}) + 2\sqrt{(2/\tau)[\log \ell(\log_2 \tau) + \log(1/\delta)]}\big) \geq 1-\delta$. We are not aware of other such generalization bounds that hold at stopping times.

Corollary 15 is comparable to the following well-known fixed-time bound which can be deduced, for instance, from the proof of Theorem 3.5 of Mohri, Rostamizadeh, and Talwalkar (2018):

$$\mathbb{P}\left(\sup_{f\in\mathcal{F}} |R_t(f) - R(f)| \leq \mathcal{R}_t(\mathcal{F}) + \sqrt{\log(2/\delta)/2t}\right) \geq 1-\delta.$$

Once again, we observe that our time-uniform bound only loses iterated logarithmic factors and small universal constants in comparison to the above display. When the population Rademacher complexity $\mathcal{R}_t(\mathcal{F})$ is unavailable in closed form, Corollary 15(ii) may be used to provide a time-uniform lower bound on this quantity in terms of its empirical counterpart. We obtain this result in Section 4.D.6 by noting that $\widehat{\mathcal{R}}_t$ satisfies the leave-one-out property in equation (4.16), although it cannot easily be written as the evaluation of a convex functional at the empirical measure. We leave open the question of providing upper confidence sequences on $\mathcal{R}_t(\mathcal{F})$, which combined with Corollary 15(i) would lead to a fully empirical bound for the classification risk.

The proof of Corollary 15(i) shows that analogous time-uniform concentration inequalities can be obtained for general suprema of empirical processes over uniformly bounded function classes, up to modifying the expectation bound $\mathcal{R}_{\bar{t}}(\mathcal{F})$, which yields time-uniform inference for the risk of arbitrary estimators with respect to a bounded loss function, in terms of their empirical risk.

4.4.7 Sequential Estimation of Multivariate Means

We next show how Theorem 15 can be used to derive confidence sequences for the mean μ of a multivariate distribution $P \in \mathcal{P}(\mathbb{R}^d)$. In the special case $d = 1$, our results show how our reverse submartingale techniques can recover known confidence sequences for univariate sequential mean estimation (summarized in Section 4.2.5), up to constant factors.

Let $(X_t)_{t=1}^{\infty}$ be a sequence of i.i.d. random variables with mean μ, and let $\mu_t = (1/t)\sum_{i=1}^t X_i$. We state our bounds in terms of a general norm $\|\cdot\|$ on $\mathbb{R}^d$, whose dual norm is denoted by $\|\cdot\|_\star = \sup_{\|\lambda\|=1}\langle \lambda, \cdot\rangle$. Assume further that there exists $\lambda_{\max} > 0$ and a convex function $\psi : [0, \lambda_{\max}) \to \mathbb{R}$ such that

$$\sup_{\nu\in\mathbb{S}_\star^{d-1}} \log\Big(\mathbb{E}_{X\sim P}\left[\exp\left(\lambda\langle \nu, X-\mu\rangle\right)\right]\Big) \le \psi(\lambda), \quad \lambda \in [0, \lambda_{\max}), \tag{4.36}$$

where $\mathbb{S}_\star^{d-1} = \{x \in \mathbb{R}^d : \|x\|_\star = 1\}$. For instance, when $\psi(\lambda) = \lambda^2\sigma^2/2$, the above definition reduces to that of a $(\sigma^2, \lambda_{\max}^{-1})$-sub-exponential random vector given in Vershynin (2018) when $\lambda_{\max} < \infty$, or of a σ^2-sub-Gaussian random vector when $\lambda_{\max} = \infty$. Finally, for any $\gamma > 0$, let N_γ denote the γ-covering number (van der Vaart and Wellner, 1996) of $\mathbb{S}_\star^{d-1}$ with respect to the norm $\|\cdot\|_\star$.

Corollary 16. *Assume P satisfies the tail assumption* (4.36)*. Then, for all $\gamma \in [0,1)$, $\delta \in (0,1)$,*

$$\mathbb{P}\left\{\exists t \ge 1 : \|\mu_t - \mu\| \ge \frac{1}{1-\gamma}(\psi^*)^{-1}\left(\frac{\log \ell(\log_2 t) + \log(1/\delta) + \log N_\gamma}{\lceil t/2\rceil}\right)\right\} \le \delta.$$

Above, $\gamma = 1/2$ is a reasonable default value. We first illustrate the result of Corollary 16 in the special case when $d = 1$ and P is 1-sub-Gaussian. If the norm $\|\cdot\|$ is taken to be the absolute value, notice that one may choose $\gamma = 0$ and $N_\gamma = 2$, thus Corollary 16 implies

$$\forall t \ge 1 : |\mu_t - \mu| \le 2\sqrt{\frac{1}{t}\Big[\log \ell(\log_2 t) + \log(2/\delta)\Big]}, \quad \text{with probability at least } 1-\delta. \tag{4.37}$$

Equation (4.37) is comparable to state-of-the-art confidence sequences for univariate means, summarized in Section 4.2.5. For example, Theorem 1 of Howard et al. (2021) provides a one-sided bound for means of 1-sub-Gaussian random variables, which together with a union bound leads to the two-sided confidence sequence

$$\forall t \ge 1 : |\mu_t - \mu| \le k_1\sqrt{\frac{1}{t}\Big[\log \ell(\log_2 t) + \log(2/\delta)\Big]}, \quad \text{with probability at least } 1-\delta, \tag{4.38}$$

where $k_1 = \frac{2^{1/4}+2^{-1/4}}{\sqrt{2}} \approx 1.8$. It can be seen from the preceding two displays that our confidence sequence is wider by a mere factor of $2/1.8 \approx 1.1$ compared to that of Howard et al.

(2021). In fact, Corollary 16 is a special case of a more general result that can be deduced from Theorem 17 in Section 4.C.1, for which there are tuning parameters that we did not optimize here, so the factor 1.1 could presumably be lowered further. More importantly though, our result applies more generally to means of multivariate distributions, for which we do not know of any other confidence sequences in the literature beyond those of Abbasi-Yadkori, Pal, and Szepesvari (2011). The latter paper only has sub-Gaussian bounds decaying at the $\sqrt{\log t/t}$ rate, instead of our $\sqrt{\log\log t/t}$ rate.

As a multivariate example, suppose now that P is (σ^2, α)-sub-exponential, so that $\psi(\lambda) = \lambda^2\sigma^2/2$ with $\lambda_{\max} = 1/\alpha$. When $\|\cdot\|$ is taken to be the Euclidean norm $\|\cdot\|_2$, one may derive the bound

$$\forall t \geq 1 : \|\mu_t - \mu\|_2 \leq 2 \begin{cases} \sqrt{2\sigma^2\gamma_t}, & t : 0 \leq \gamma_t < \frac{\sigma^2}{2\alpha^2}, \\ \gamma_t\alpha + \frac{\sigma^2}{2\alpha}, & t : \frac{\sigma^2}{2\alpha^2} \leq \gamma_t, \end{cases}, \quad \text{with probability at least } 1-\delta,$$

where $\gamma_t = (2/t)[\log \ell(\log_2 t) + \log(1/\delta) + d\log 5]$. In particular, we recover the optimal dependence on both d and t from the fixed-time setting, up to iterated logarithmic factors.

4.4.8 An Upper Law of the Iterated Logarithm for sub-Gaussian Divergences

We now show how our finite-sample results can be used to derive an asymptotic statement which mirrors the classical (upper) law of the iterated logarithm (LIL; Stout (1970)) for sums of i.i.d. random variables.

Corollary 17. *Let D be a convex divergence such that $D(P_t\|P)$ is (σ^2/t)-sub-Gaussian for all $t \geq 1$ and some $\sigma > 0$. Assume $\mathbb{E}D(P_t\|P) = o(\sqrt{(\log\log t)/t})$. Then,*

$$\limsup_{t\to\infty} \frac{tD(P_t\|P)}{\sqrt{2t\sigma^2\log\log t}} \leq 1, \quad \textit{a.s.}$$

Corollary 17 establishes an upper LIL for convex divergences admitting the same constant as the classical LIL, which states that for any sequence of mean-zero i.i.d. random variables $(X_t)_{t=1}^\infty$ admitting finite variance $\sigma^2 > 0$,

$$\limsup_{t\to\infty} \frac{1}{\sqrt{2t\sigma^2\log\log t}} \sum_{i=1}^t X_i = 1, \quad \text{a.s.}$$

Obtaining a matching lower bound in Corollary 17 would, for instance, necessitate anti-concentration bounds on the process $D(P_t\|P)$, and is therefore beyond the scope of this work. The sub-Gaussianity assumption can also likely be weakened, but given again that our purpose was not asymptotics, we leave this for future work. Though results analogous to Corollary 17 have possibly appeared in past literature for various divergences, we are only aware of the LILs for the Kolmogorov-Smirnov statistic derived by Smirnov (1944), for the 1-Wasserstein distance in dimension $d = 1$ (del Barrio et al., 1999), and for certain von Mises differentiable functionals (Serfling, 2009).

Adaptations of the proofs of Corollaries 9, 11, 13, and 20 respectively imply that Corollary 17 holds when D is taken to be the Kolmogorov-Smirnov distance, the Maximum Mean Discrepancy with bounded kernel, the Total Variation distance for distributions supported on a finite set, or the Total Variation and 1-Wasserstein distances smoothed by Gaussian convolution under suitable tail assumptions on P. The conditions of Corollary 17 can also be verified for the transportation cost W_p^p, for any $p \geq 1$ satisfying $p > d/2$, under suitable moment assumptions on P (Fournier and Guillin, 2015; Niles-Weed and Rigollet, 2022). To the best of our knowledge, this functional is not known to be von Mises differentiable without further assumptions on the connectedness of the support or absolute continuity of P (see Goldfeld et al. (2024) and Chapter 7 below).

4.A Additional Lemmas

We begin by recalling McDiarmid's inequality (Wainwright, 2019). We say that a map $G : \mathbb{R}^t \to \mathbb{R}$ satisfies the bounded differences property with parameters $(L_1, \dots, L_t) \in \mathbb{R}_+^t$ if for every $x_1, \dots, x_t, x_1', \dots, x_t' \in \mathbb{R}$ and all $i = 1, \dots, t$,

$$|G(x_1, \dots, x_t) - G(x_1, \dots, x_{i-1}, x_i', x_{i+1}, \dots, x_t)| \leq L_i.$$

Theorem 16 (McDiarmid's Inequality). *Assume G satisfies the bounded differences property with parameters $(L_1, \dots, L_t)$ and that $X_1, \dots, X_t$ are independent random variables. Then, for all $u \geq 0$,*

$$\mathbb{P}\Big(\big|G(X_1, \dots, X_t) - \mathbb{E}G(X_1, \dots, X_t)\big| \geq u\Big) \leq 2\exp\left\{-\frac{2u^2}{\sum_{i=1}^t L_i^2}\right\}.$$

In several of the proofs for Section 4.4, we will show that the processes $\Phi(P_t, Q)$ or $\Psi(P_t, Q_s)$ satisfy the bounded differences property, when viewed as functions of the samples therein. The following standard result will then imply that these processes are sub-Gaussian (see for instance Rigollet and Hütter (2015), Lemma 1.5, for a statement with the exact constants used below).

Lemma 32. *Let P be a distribution over $\mathbb{R}$ such that for $X \sim P$ and $\sigma > 0$, $\mathbb{E}[X] = 0$ and*

$$\mathbb{P}(|X| > u) \leq 2e^{-u^2/2\sigma^2}, \quad u > 0.$$

Then, P is $(8\sigma^2)$-sub-Gaussian.

4.B Proofs from Section 4.2

4.B.1 Proofs of Theorem 13

Theorem 13 was proven for instance by Lee (1990), Theorem 3, p. 112. Due to its centrality in our work, we provide two self-contained proofs of this result below. The first proof follows directly from Doob's submartingale inequality (see equation (4.10)). The second proof is a restatement of Lee's original proof, which we include for reference in the following subsection.

First Proof (via Doob's submartingale inequality). For any integer $T \geq t_0$, define the process $S_t = R_{T-t+t_0}$, for all $t_0 \leq t \leq T$, as well as the forward filtration $\mathcal{G}_t = \mathcal{F}_{T-t+t_0}$. Since (R_t) is a reverse submartingale, we have $\mathbb{E}[R_t|\mathcal{F}_{t+1}] \geq R_{t+1}$, whence for all $t_0+1 \leq t \leq T$,

$$\mathbb{E}[S_t|\mathcal{G}_{t-1}] = \mathbb{E}[R_{T-t+t_0}|\mathcal{F}_{T-t+t_0+1}] \geq R_{T-t+t_0+1} = S_{t-1}.$$

It follows that $(S_t)_{t=t_0}^T$ forms a forward submartingale with respect to $(\mathcal{G}_t)_{t=t_0}^T$. Applying Doob's submartingale inequality, we therefore obtain

$$\mathbb{P}(\exists t_0 \leq t \leq T : S_t \geq u) \leq \frac{\mathbb{E}[S_T]}{u},$$

for all $u > 0$. Equivalently,

$$\mathbb{P}(\exists t_0 \leq s \leq T : R_s \geq u) \leq \frac{\mathbb{E}[R_{t_0}]}{u}.$$

Notice that the event within the probability on the left-hand side of the above display is monotonically increasing with T, converging to the event $\{\exists s \geq t_0 : R_s \geq u\}$. Taking $T \to \infty$, we thus have

$$\mathbb{P}(\exists s \geq t_0 : R_s \geq u) \leq \frac{\mathbb{E}[R_{t_0}]}{u},$$

which proves the claim. □

Second Proof (first principles). Let $T \geq t_0$, $u > 0$, and define the disjoint sets

$$A_t = \{R_t \geq u\} \cap \bigcap_{j=t+1}^{T} \{R_j < u\}, \quad t = t_0, t_0+1, \dots, T,$$

where t represents the last time (in the range $t_0, t_0+1, \dots, T$) at which R_t was larger than u. We have,

$$\mathbb{P}(\exists t_0 \leq t \leq T : R_t \geq u) = \sum_{t=t_0}^{T} \mathbb{P}(A_t) \leq \frac{1}{u}\sum_{t=t_0}^{T}\int_{A_t} R_t d\mathbb{P} \leq \frac{1}{u}\sum_{t=t_0}^{T}\int_{A_t} \mathbb{E}[R_{t_0}|\mathcal{F}_t] d\mathbb{P},$$

where the first inequality follows because $R_t > u$ on A_t, and the second inequality follows by the reverse submartingale property of (R_t). Note that R_j is $\mathcal{F}_j$-measurable and hence also $\mathcal{F}_t$-measurable for $t \leq j$ due to the reversed nature of the filtrations. Thus, $A_t \in \mathcal{F}_t$, whence $\int_{A_t} \mathbb{E}[R_{t_0}|\mathcal{F}_t] d\mathbb{P} = \int_{A_t} R_{t_0} d\mathbb{P}$ and we obtain

$$\mathbb{P}(\exists t_0 \leq t \leq T : R_t \geq u) \leq \frac{1}{u}\sum_{t=t_0}^{T}\int_{A_t} R_{t_0} d\mathbb{P} \leq \frac{1}{u}\mathbb{E}[R_{t_0}],$$

where the last step utilizes the nonnegativity of R_t and the fact that the events $\{A_t\}_{t=t_0}^T$ are disjoint by construction. The claim now follows as before by taking $T \to \infty$, noting that the right-hand side remains fixed while the left-hand side is the probability of an increasing sequence of events whose limit is $\{\exists t \geq t_0 : R_t \geq u\}$. □

4.B.2 Proof of Proposition 13

Let $(\mathcal{F}_{ts})_{t,s\geq 1}$ be a reverse filtration. Under the conditional independence condition (4.12) on $(\mathcal{F}_{ts})$, Christofides and Serfling (1990) establish a maximal inequality for partially ordered reverse submartingales, which we state below before proving Proposition 13. We will require the following notation. Let $1 \leq t_0 \leq T, 1 \leq s_0 \leq S$, and let $\{C_{ts} : t, s \geq 0\}$ be a nondecreasing array of nonnegative numbers. Given a reverse submartingale $(R_{ts})_{t,s\geq 1}$ with respect to $(\mathcal{F}_{ts})$, a general bound will be given on $\mathbb{P}(A)$, where for any $u > 0$,

$$\begin{aligned} A = \left\{ \max_{\substack{t_0\leq t\leq T \\ s_0\leq s\leq S}} C_{ts}R_{ts} \geq u \right\} &= \bigcup_{t=t_0}^{T} \bigcup_{s=s_0}^{S} A_{ts}, \\ \text{where } \; A_{ts} = \{R_{ts}C_{ts} \geq u\} &\cap \bigcap_{j=t+1}^{T} \bigcap_{k=s+1}^{S} \{R_{jk}C_{jk} < u\}. \end{aligned} \tag{4.39}$$

This decomposition into the sets A_{ts} is analogous to that in the proof of Ville's inequality for reverse submartingales (Section 4.B.1). Unlike that result, however, the lack of total ordering on $\mathbb{N}^2$ prevents the sets A_{ts} from being disjoint; for instance A_{43} and A_{34} could both potentially happen. Christofides and Serfling (1990) instead form a partition $(B_{ts}^{(1)})_{t,s\geq 1}$ of A, defined recursively by the following algorithm.

```
Let $D_0 = \emptyset,\ m := 1$
    For $j = t_0$ to $T$
        For $k = s_0$ to $S$
            $B_{jk}^{(1)} := A_{jk} \setminus \bigcup_{l<m} D_l$
            $D_m := A_{jk},\ m := m + 1.$
```

A second partition $(B_{ts}^{(2)})_{t,s\geq 1}$ is further formed by changing the order of the for-loops in the above display. Specifically,

$$\begin{aligned} B_{ts}^{(1)} &= A_{ts} \setminus \left\{ \left(\bigcup_{j=t_0}^{t-1} \bigcup_{k=s_0}^{S} A_{jk} \right) \cup \left(\bigcup_{k=s_0}^{s-1} A_{tk} \right) \right\}, \\ B_{ts}^{(2)} &= A_{ts} \setminus \left\{ \left(\bigcup_{k=s_0}^{s-1} \bigcup_{j=t_0}^{T} A_{jk} \right) \cup \left(\bigcup_{j=t_0}^{t-1} A_{js} \right) \right\}, \end{aligned}$$

with the convention that an empty union is equal to the empty set. Notice that for $j = 1, 2$, the sets $(B_{ts}^{(j)})_{t,s\geq 1}$ are mutually disjoint, and $\bigcup_{t,s\geq 1} B_{ts}^{(j)} = A$. Further, unlike (A_{ts}), the sequence $(B_{ts}^{(j)})$ is not adapted to $(\mathcal{F}_{ts})$, but instead satisfies $B_{ts}^{(1)} \in \mathcal{F}_{ts_0}$ and $B_{ts}^{(2)} \in \mathcal{F}_{t_0 s}$. We are now in a position to state their bound.

Lemma 33 (Christofides and Serfling (1990), Corollary 2.9)**.** *Let (R_{ts}) be a nonnegative reverse submartingale with respect to $(\mathcal{F}_{ts})$, and assume $(\mathcal{F}_{ts})$ satisfies the conditional independence condition* (4.12)*. Furthermore, let $\{C_{ts} : t, s \geq 0\}$ be a nondecreasing array of nonnegative numbers. Then, for all $u > 0$,*

$$\begin{aligned} &u\mathbb{P}\left\{\max_{\substack{t_0\leq t\leq T\\ s_0\leq s\leq S}} C_{ts}R_{ts} \geq u\right\} \\ &\quad\leq \left\{\sum_{t=t_0}^{T}\sum_{s=s_0}^{S}(C_{ts}-C_{(t-1)s})\mathbb{E}[R_{ts}] - \sum_{s=s_0}^{S} C_{t_0 s}\int_{(\bigcup_{t=t_0}^{T} B_{ts}^{(1)})^c} R_{t_0 s}d\mathbb{P}\right\} \\ &\quad\wedge \left\{\sum_{t=t_0}^{T}\sum_{s=s_0}^{S}(C_{ts}-C_{t(s-1)})\mathbb{E}[R_{ts}] - \sum_{t=t_0}^{T} C_{t s_0}\int_{(\bigcup_{s=s_0}^{S} B_{ts}^{(2)})^c} R_{t s_0}d\mathbb{P}\right\}. \end{aligned}$$

In the special case where $C_{ts} = I(t \geq t_0, s \geq s_0)$ for all $0 \leq t \leq T, 0 \leq s \leq S$, Lemma 33 reduces to the following bound

$$\begin{aligned} \mathbb{P}\left\{\max_{\substack{t_0\leq t\leq T\\ s_0\leq s\leq S}} R_{ts} \geq u\right\} &\leq \frac{1}{u}\left[\sum_{s=s_0}^{S} C_{t_0 s}\mathbb{E}[R_{t_0 s}] - \sum_{s=s_0}^{S} C_{t_0 s}\int_{(\bigcup_{t=t_0}^{T} B_{ts}^{(1)})^c} R_{t_0 s}d\mathbb{P}\right] \\ &= \frac{1}{u}\sum_{s=s_0}^{S}\int_{\bigcup_{t=t_0}^{T} B_{ts}^{(1)}} R_{t_0 s}d\mathbb{P}. \end{aligned} \tag{4.40}$$

This simplification of Lemma 33 turns out to be simple to show, and we provide a self-contained proof before using it to prove Proposition 13 below.

Proof of Inequality (4.40)**.** We have for any $s_0 \leq s \leq S$,

$$\begin{aligned} u\mathbb{P}\left(\bigcup_{t=t_0}^{T} B_{ts}^{(1)}\right) &= u\sum_{t=t_0}^{T}\mathbb{P}(B_{ts}^{(1)}) \\ &\leq \sum_{t=t_0}^{T}\int_{B_{ts}^{(1)}} R_{ts}d\mathbb{P} \\ &\leq \sum_{t=t_0}^{T}\int_{B_{ts}^{(1)}} \mathbb{E}[R_{t_0 s}|\mathcal{F}_{ts}]d\mathbb{P} \\ &= \sum_{t=t_0}^{T}\int_{B_{ts}^{(1)}} \mathbb{E}\Big\{\mathbb{E}[R_{t_0 s}|\mathcal{F}_{t_0 s}] \mid \mathcal{F}_{ts_0}\Big\}d\mathbb{P} && \text{(By the CI condition)} \\ &= \sum_{t=t_0}^{T}\int_{B_{ts}^{(1)}} \mathbb{E}[R_{t_0 s}|\mathcal{F}_{t_0 s}]d\mathbb{P} && \text{(Since } B_{ts}^{(1)} \in \mathcal{F}_{ts_0}\text{)} \\ &= \sum_{t=t_0}^{T}\int_{B_{ts}^{(1)}} R_{t_0 s}d\mathbb{P} = \int_{\bigcup_{t=t_0}^{T} B_{ts}^{(1)}} R_{t_0 s}d\mathbb{P}. \end{aligned}$$

The claim follows by taking a summation over $s_0 \le s \le S$ on both sides.

Lemma 33 leads to the following proof of Proposition 13, which generalizes Corollary 2.10 of Christofides and Serfling (1990).

Proof of Proposition 13. Let $T \ge t_0, S \ge s_0$. By inequality (4.40),

$$\begin{aligned} u^\alpha \mathbb{P}\left\{ \sup_{\substack{t_0 \le t \le T \\ s_0 \le s \le S}} R_{ts} \ge u \right\} &\le \sum_{s=s_0}^{S} \int_{\bigcup_{t=t_0}^{T} B_{ts}^{(1)}} R_{t_0 s}^\alpha d\mathbb{P} \\ &\le \sum_{s=s_0}^{S} \int_{\bigcup_{t=t_0}^{T} B_{ts}^{(1)}} \left(\max_{s_0 \le s \le S} R_{t_0 s}^\alpha \right) d\mathbb{P} \\ &= \int_A \left(\max_{s_0 \le s \le S} R_{t_0 s}^\alpha \right) d\mathbb{P} \\ &\le \mathbb{E}\left(\max_{s_0 \le s \le S} R_{t_0 s}^\alpha \right) \\ &\le \left(\frac{\alpha}{\alpha - 1} \right)^\alpha \mathbb{E}[R_{t_0 s_0}^\alpha], \end{aligned}$$

where the last inequality follows from Doob (1953), Theorem 3.4, page 317. Taking $T, S \to \infty$ on both sides of the above display leads to the claim. □

4.C Proofs from Section 4.3

4.C.1 Proofs from Subsection 4.3.1

Proof of Proposition 15. To prove part (i), Theorem 14 implies that (N_t) forms a reverse submartingale with respect to $(\mathcal{E}_t^X)$. Furthermore, the map $x \in \mathbb{R} \mapsto \exp\{\lambda x\}$ is convex and monotonic for any fixed $\lambda \in [0, \lambda_{\max})$, so Jensen's inequality implies that the process

$$L_t(\lambda) = \exp(\lambda N_t), \quad t \ge 1,$$

is also a reverse submartingale with respect to $(\mathcal{E}_t^X)$. By Theorem 13, we obtain for all $u > 0$,

$$\begin{aligned} \mathbb{P}\Big(\exists t \ge t_0 : N_t \ge u\Big) &\le \inf_{\lambda \in [0, \lambda_{\max})} \mathbb{P}\Big(\exists t \ge t_0 : L_t(\lambda) \ge e^{\lambda u}\Big) \le \inf_{\lambda \in [0, \lambda_{\max})} \mathbb{E}\left[\exp(-\lambda u) L_{t_0}(\lambda)\right] \\ &\le \inf_{\lambda \in [0, \lambda_{\max})} \exp\big\{ -\lambda u + \psi_{t_0}(\lambda)\big\} = \exp\big\{ -\psi_{t_0}^*(u)\big\}, \end{aligned} \tag{4.41}$$

as claimed. To prove Proposition 15(ii), recall that $\mathcal{E}_{ts} = \mathcal{E}_t^X \bigvee \mathcal{E}_s^Y$ is a σ-algebra generated by a union of independent σ-algebras. It follows that the filtration $(\mathcal{E}_{ts})$ satisfies the conditional independence property (4.12), by Cairoli and Walsh (1975), example (a), page 114 (see also Christofides and Serfling (1990)). Furthermore, similarly as in part (i), the process

$$L_{ts}(\lambda, \alpha) = \exp(\lambda M_{ts}/\alpha), \quad t, s \ge 1,$$

is a partially ordered reverse submartingale for any fixed choice of $\lambda \in [0, \lambda_{\max})$ and $\alpha > 1$. Notice also that $L_{ts}(\lambda, \alpha) \in L^\alpha(\mathbb{P})$, thus we may apply Proposition 13 to obtain

$$\begin{aligned}\mathbb{P}(\exists t \geq t_0, s \geq s_0 : M_{ts} \geq u) &= \inf_{\lambda \in [0,\lambda_{\max})} \mathbb{P}\Big(\exists t \geq t_0, s \geq s_0 : L_{ts}(\lambda, \alpha) \geq \exp(\lambda u/\alpha)\Big) \\ &\leq \left(\frac{\alpha}{\alpha-1}\right)^\alpha \inf_{\lambda \in [0,\lambda_{\max})} \exp(-\lambda u)\mathbb{E}\left[L^\alpha_{t_0 s_0}(\lambda, \alpha)\right] \\ &= \left(\frac{\alpha}{\alpha-1}\right)^\alpha \inf_{\lambda \in [0,\lambda_{\max})} \exp(-\lambda u)\mathbb{E}\left[\exp(\lambda M_{t_0 s_0})\right] \\ &\leq \left(\frac{\alpha}{\alpha-1}\right)^\alpha \exp(-\psi^*_{t_0 s_0}(u)).\end{aligned}$$

Taking the infimum over $\alpha > 1$ on both sides of the above display leads to the claim. □

We now turn to proving a more general version of Theorem 15. Let $\eta, \xi > 1$ be fixed constants which determine the sizes of the geometric epochs used in the proofs. Furthermore, given $t, s \geq 1$, we use the shorthand notation $\bar{t} = \lceil t/\lceil \eta \rceil \rceil$ and $\bar{s} = \lceil s/\lceil \xi \rceil \rceil$.

Theorem 17. *Let Φ, Ψ be convex functionals, and let $\delta \in (0, 1)$.*

(i) (One-Sample) Assume ψ^*_t is invertible for all $t \geq 1$, and if η is not an integer, assume $(\psi^*_t)^{-1}(\lambda)$ is a decreasing (resp. increasing) sequence in t (resp. λ). Set

$$\gamma_t = (\psi^*_{\bar{t}})^{-1}\Big(\log \ell(\log_\eta t) + \log(1/\delta)\Big).$$

Assume further that (γ_t) is a nonincreasing sequence. Then,

$$\mathbb{P}\Big\{\exists t \geq 1 : \Phi(P_t) \geq \Phi(P) + \gamma_t\Big\} \leq \delta.$$

(ii) (Two-Sample) Assume ψ^*_{ts} is invertible for all $t, s \geq 1$, and if η (resp ξ) is not an integer, assume $\psi^*_{ts}(\lambda)$ is decreasing in t (resp. in s), and increasing in λ. Set

$$\gamma_{ts} = (\psi^*_{\bar{t}\bar{s}})^{-1}\Big(\log g(\log_\eta t + \log_\xi s) + \log(1/\delta)\Big).$$

Assume further that (γ_{ts}) is a nonincreasing sequence in each of its indices. Then,

$$\mathbb{P}\left\{\exists t, s \geq 1 : \Psi(P_t, Q_s) \geq \Psi(P, Q) + \gamma_{ts}\right\} \leq \delta.$$

Proof of Theorems 15 and 17. The proofs of claims (i) and (ii) of Theorem 17 are similar, thus we only prove (ii). Theorem 15 will then follow by setting $\eta = \xi = 2$. Let $u_j = \lceil \eta^j \rceil$ and $v_k = \lceil \xi^k \rceil$ for all $j, k \in \mathbb{N}_0$. Since γ_{ts} is decreasing in t and s, we have

$$\begin{aligned}&\mathbb{P}\left(\exists t, s \geq 1 : M_{ts} \geq \gamma_{ts}\right) \\ &\leq \mathbb{P}\left(\bigcup_{j \in \mathbb{N}_0} \bigcup_{k \in \mathbb{N}_0} \Big\{\exists t \in \{u_j, \dots, u_{j+1}\}, s \in \{v_k, \dots, v_{k+1}\} : M_{ts} \geq \gamma_{ts}\Big\}\right)\end{aligned}$$

$$\leq \mathbb{P}\left(\bigcup_{j\in\mathbb{N}_0}\bigcup_{k\in\mathbb{N}_0}\Big\{\exists t \in \{u_j,\dots,u_{j+1}\}, s\in\{v_k,\dots,v_{k+1}\} : M_{ts} \geq \gamma_{u_{j+1}v_{k+1}}\Big\}\right).$$

Now, $u_{j+1} \leq u_j\lceil\eta\rceil$ (resp. $v_{k+1} \leq v_k\lceil\xi\rceil$), with equality if η (resp. ξ) is an integer. Therefore, by definition of $\bar{t}, \bar{s}$, and by the fact that $(\psi^*_{ts})^{-1}$ is decreasing in t (resp. s) when η (resp. ξ) is not an integer, we have

$$\gamma_{u_{j+1}v_{k+1}} \geq (\psi^*_{u_jv_k})^{-1}\Big(\log g\big(\log_\eta(u_{j+1}) + \log_\xi(v_{k+1})\big) + \log(1/\delta)\Big).$$

Since $(\psi^*_{ts})^{-1}(\lambda)$ is increasing in λ when η or ξ are not integers, we deduce

$$\gamma_{u_{j+1}v_{k+1}} \geq (\psi^*_{u_jv_k})^{-1}\Big(\log g(j+k+2) + \log(1/\delta)\Big).$$

Applying a union bound together with Proposition 15 then leads to

$$\begin{aligned}
&\mathbb{P}\left(\exists t, s \geq 1 : M_{ts} \geq \gamma_{ts}\right)\\
&\leq e\sum_{j=0}^\infty\sum_{k=0}^\infty \exp\Big\{-\psi^*_{u_jv_k}\Big((\psi^*_{u_jv_k})^{-1}\big(\log g(j+k+2)+\log(1/\delta)\big)\Big)\Big\}\\
&\leq e\sum_{j=0}^\infty\sum_{k=0}^\infty \exp\big\{-\big(\log g(j+k+2)+\log(1/\delta)\big)\big\}\\
&= \delta\sum_{j=0}^\infty\sum_{k=0}^\infty \frac{e}{g(j+k+2)} \leq \delta,
\end{aligned}$$

as claimed. □

Proof of Corollary 6. We only prove the claim for $(M_{ts})_{t,s=1}^\infty$, as the proof for $(N_t)_{t=1}^\infty$ is similar. Since M_{ts} is σ^2_{ts}-sub-Gaussian, we have for all $\lambda \in \mathbb{R}_+$,

$$\begin{aligned}
\mathbb{E}\left\{\exp\left(\lambda M_{ts}\right)\right\} &= \mathbb{E}\left\{\exp\left[\lambda(M_{ts} - \mathbb{E}(M_{ts}))\right]\right\}\exp\left\{\lambda\mathbb{E}(M_{ts})\right\}\\
&\leq \exp\left\{\frac{\lambda^2\sigma^2_{ts}}{2}\right\}\exp\left\{\lambda\mathbb{E}(M_{ts})\right\},
\end{aligned}$$

whence, an upper bound on the CGF of M_{ts} is given by $\psi_{ts}(\lambda) = \frac{\lambda^2\sigma^2_{ts}}{2} + \lambda\mathbb{E}(M_{ts})$. Thus, for any $x \geq \mathbb{E}(M_{ts})$ and any $\gamma \in \mathbb{R}_+$,

$$\psi^*_{ts}(x) = \frac{(x-\mathbb{E}(M_{ts}))^2\sigma^{-2}_{ts}}{2}, \quad (\psi^*_{ts})^{-1}(\gamma) = \mathbb{E}(M_{ts}) + \sqrt{2\gamma\sigma^2_{ts}}.$$

The claim now follows from Theorem 15. □

4.C.2 Proofs from Subsection 4.3.2

We state and prove the following stronger version of Corollary 8.

Proposition 21. Assume that the processes (N_t) and (M_{ts}) satisfy the conditions of Proposition 16, and fix $\delta \in (0,1)$. Assume further that there exists $\lambda_{\max} > 0$ and convex functions $\psi_\Phi, \psi_\Psi, \phi_\Psi$ such that for all $\lambda \in [0, \lambda_{\max})$,

$$\sup_{f \in \mathcal{F}_\Phi} \log \left\{ \mathbb{E}\left[e^{\lambda(f(X) - \mathbb{E}f(X))} \right] \right\} \leq \psi_\Phi(\lambda),$$

$$\sup_{f \in \mathcal{F}_\Psi} \log \left\{ \mathbb{E}\left[e^{\lambda(f(X) - \mathbb{E}f(X))} \right] \right\} \leq \psi_\Psi(\lambda), \quad \sup_{g \in \mathcal{G}_\Psi} \log \left\{ \mathbb{E}\left[e^{\lambda(g(Y) - \mathbb{E}g(Y))} \right] \right\} \leq \phi_\Psi(\lambda).$$

Assume further that the Legendre-Fenchel transforms $\psi_\Phi^*, \psi_\Psi^*, \phi_\Psi^*$ are invertible, with nondecreasing inverses.

(i) (One-Sample) We have

$$\mathbb{P}\left\{ \exists t \geq 1 : \Phi(P_t) \leq \Phi(P) - (\psi_\Phi^*)^{-1}\left(\frac{2}{t}\Big[\log \ell(\log_2 t) + \log(2/\delta) \Big] \right) \right\} \leq \delta/2.$$

(ii) (Two-Sample) Define

$$\begin{aligned} \kappa_t^X &= (\psi_\Psi^*)^{-1}\left(\frac{2}{t}\left[\log \ell(\log_2 t) + \log\left(\frac{4}{\delta}\right) \right] \right), \\ \kappa_s^Y &= (\phi_\Psi^*)^{-1}\left(\frac{2}{s}\left[\log \ell(\log_2 s) + \log\left(\frac{4}{\delta}\right) \right] \right). \end{aligned}$$

Then, we have

$$\mathbb{P}\big(\exists t, s \geq 1 : \Psi(P_t, Q_s) \leq \Psi(P, Q) - \kappa_t^X - \kappa_s^Y \big) \leq \delta/2. \tag{4.42}$$

When it exists, the cumulant generating function of any mean-zero random variable Z scales quadratically near zero—specifically, it is easy to check that $\lim_{\lambda \to 0} \log(\mathbb{E}\exp(\lambda Z))/(\lambda^2/2) = \mathrm{Var}(Z)$. Thus, the inverse of its Legendre-Fenchel transform typically scales as the square root function near zero. The upper confidence sequences in Proposition 21 thus typically scale at the parametric rate up to a necessary iterated logarithmic factor.

Proof of Proposition 21. To prove part (i), define for this proof only,

$$\eta_t = (\psi_\Phi^*)^{-1}\left(\frac{2}{t}\Big[\log \ell(\log_2 t) + \log(2/\delta) \Big] \right).$$

Furthermore, let $u_j = 2^j$ for all $j \in \mathbb{N}_0$. By Proposition 16, (R_t) minorizes (N_t) thus

$$\mathbb{P}(\exists t \geq 1 : \Phi(P_t) \leq \Phi(P) - \eta_t) \leq \mathbb{P}\left(\exists t \geq 1 : -R_t \geq \eta_t \right)$$

$$\leq \mathbb{P}\left(\bigcup_{j\in\mathbb{N}_0}\Big\{\exists t \in \{u_j,\dots,u_{j+1}\} : -R_t \geq \eta_t\Big\}\right)$$
$$\leq \mathbb{P}\left(\bigcup_{j\in\mathbb{N}_0}\left\{\exists t \in \{u_j,\dots,u_{j+1}\} : -R_t \geq (\psi_\Phi^*)^{-1}\left(\frac{2}{u_{j+1}}\left[\log \ell(j+1) + \log(2/\delta)\right]\right)\right\}\right), \tag{4.43}$$

where on the last line, we used the fact that $(\psi_\Phi^*)^{-1}$ is nondecreasing. Now, (R_t) is a reverse martingale with respect to $(\mathcal{E}_t^X)$, whence $(\exp(-\lambda R_t))_{t=1}^\infty$ is a reverse submartingale for any fixed $\lambda \in [0, t_0\lambda_{\max})$. Applying Theorem 13 similarly as in the proof of Proposition 15, we therefore obtain for all $u > 0$ and $t_0 \geq 1$,

$$\begin{aligned}\mathbb{P}\Big(\exists t \geq t_0 : -R_t \geq u\Big) &= \inf_{\lambda\in[0,t_0\lambda_{\max})} \mathbb{P}\Big(\exists t \geq t_0 : \exp(-\lambda R_t) \geq e^{\lambda u}\Big)\\ &\leq \inf_{\lambda\in[0,t_0\lambda_{\max})} \exp(-\lambda u)\mathbb{E}\left[\exp(-\lambda R_{t_0})\right]\\ &\leq \inf_{\lambda\in[0,t_0\lambda_{\max})} \exp\big\{-\lambda u + t_0\psi_\Phi(\lambda/t_0)\big\}\\ &\leq \inf_{\lambda\in[0,t_0\lambda_{\max})} \exp\big\{-t_0[(\lambda/t_0)u - \psi_\Phi(\lambda/t_0)]\big\}\\ &= \inf_{\lambda\in[0,\lambda_{\max})} \exp\big\{-t_0[\lambda u - \psi_\Phi(\lambda)]\big\} = \exp\big\{-t_0\psi_\Phi^*(u)\big\}.\end{aligned}$$

Returning to equation (4.43), we deduce

$$\mathbb{P}(\exists t \geq 1 : \Phi(P_t) \leq \Phi(P) - \eta_t) \leq \sum_{j=0}^\infty \exp\left\{-\frac{2u_j}{u_{j+1}}\left[\log \ell(j+1) + \log(2/\delta)\right]\right\}$$
$$\leq \sum_{j=0}^\infty \exp\left\{-\left[\log \ell(j+1) + \log(2/\delta)\right]\right\} = \sum_{j=0}^\infty \frac{\delta/2}{\ell(j+1)} \leq \frac{\delta}{2}.$$

The proof of claim (i) follows. The proof of part (ii) follows by a similar probability bound for each of $-R_t^X$ and $-R_s^Y$ at level $\delta/4$, combined with a union bound. □

Proof of Corollary 8. By Hoeffding's Lemma, we may take $\psi_\Phi(\lambda) = \lambda^2 B^2/8$ for all $\lambda \in \mathbb{R}_+$, thus, $(\psi_\Phi^*)^{-1}(\lambda) = \sqrt{B^2\lambda/2}$, and similarly for the two-sample case. The claim follows directly from Proposition 21. □

4.C.3 Proofs from Subsection 4.3.3

Proof of Proposition 17. Assume first that $\mathbb{P}(\bigcup_{t,s=1}^\infty A_{ts}) \leq \delta$. Let $\eta = (T, S)$ be any random time. Then

$$A_{TS} = \left(\bigcup_{t=1}^\infty \bigcup_{s=1}^\infty A_{ts} \cap \{\eta = (t,s)\}\right)$$

$$\cup \left(\bigcup_{t=1}^{\infty} A_{t\infty} \cap \{\eta = (t,\infty)\}\right)$$
$$\cup \left(\bigcup_{s=1}^{\infty} A_{\infty s} \cap \{\eta = (\infty,s)\}\right) \cup \left(A_{\infty\infty} \cap \{\eta = (\infty,\infty)\}\right).$$

Since $A_{\infty\infty}, A_{t\infty}, A_{\infty s} \subseteq \bigcup_{t,s=1}^{\infty} A_{ts}$, we deduce that $A_{TS} \subseteq \bigcup_{t,s=1}^{\infty} A_{ts}$, implying that $\mathbb{P}(A_{TS}) \leq \delta$. Thus (i) implies (iii). It is also clear that (iii) implies (ii), thus it remains to prove that (ii) implies (i). To this end, assume $\mathbb{P}(A_{\tau\sigma}) \leq \delta$ for any stopping time (τ,σ). For any $\omega \in \Omega$, let

$$I(\omega) = \left\{(t,s) \in \mathbb{N}^2 : \omega \in A_{ts} \text{ and } \omega \notin A_{t's'}, \forall (t',s') \in \mathbb{N}^2, (t',s') < (t,s)\right\}.$$

We may then define

$$(\tau(\omega),\sigma(\omega)) = \begin{cases} (\infty,\infty), & I(\omega) = \emptyset \\ \operatorname*{argmin}_{(t,s)\in I(\omega)} t, & |I(\omega)| \geq 1 \end{cases}.$$

The minimizer in the above display is unique and unambiguous, because when $I(\omega)$ has cardinality greater or equal to 2, any of its distinct elements (t,s) and (t',s') must have $t \neq t'$ and $s \neq s'$ by construction; for example, (t,s) and (t,s') cannot both be elements of $I(\omega)$ for $s \neq s'$. Notice that (τ,σ) is a stopping time with respect to $(\mathcal{F}_{ts})$ because $A_{t's'} \in \mathcal{F}_{ts}$ for all $(t',s') \leq (t,s)$. Furthermore, its definition guarantees $\bigcup_{t,s=1}^{\infty} A_{ts} \subseteq A_{\tau\sigma}$. We deduce by assumption that $\mathbb{P}(\bigcup_{t,s=1}^{\infty} A_{ts}) \leq \delta$, as claimed. □

We note that our precise definitions of the events $A_{t\infty}$ and $A_{\infty s}$, for $t,s \geq 1$, was not crucial in the preceding argument.

4.D Proofs from Section 4.4

4.D.1 Proofs from Subsection 4.4.1

Proof of Corollary 9. By the DKW inequality, for all $u > 0$, we have

$$\mathbb{P}\Big(\|F_t - F\|_\infty \geq u\Big) \leq 2e^{-2tu^2}.$$

Therefore,

$$\mathbb{E}[\|F_t - F\|_\infty] = \int_0^\infty \mathbb{P}(\|F_t - F\|_\infty \geq u)du \leq 2\int_0^\infty e^{-2tu^2}du = \sqrt{\frac{\pi}{2t}}.$$

Furthermore, it is a straightforward observation that the map

$$(x_1,\dots,x_t) \in \mathbb{R}^t \mapsto \sup_{x\in\mathbb{R}} \left| \frac{1}{t}\sum_{i=1}^t I(x_i \leq x) - F(x)\right|$$

satisfies the bounded differences property with parameters $L_1 = \cdots = L_t = 1/t$. McDiarmid's inequality (Theorem 16) therefore implies the bound

$$\mathbb{P}\big(\big|\, \|F_t - F\|_\infty - \mathbb{E}\, \|F_t - F\|_\infty \,\big| \geq u\big) \leq 2e^{-2tu^2}, \quad u > 0.$$

Lemma 32 then implies that $\|F_t - F\|_\infty$ is a $(2/t)$-sub-Gaussian random variable.

Now, notice that $D_{\mathcal{J}}(P_t\|P) = \|F_t - F\|_\infty$ is an IPM over the class $\mathcal{J} = \{(-\infty, x] : x \in \mathbb{R}\}$, and in particular $D_{\mathcal{J}}$ is convex by Lemma 31. We may therefore apply Corollary 6 to the process $N_t = D_{\mathcal{J}}(P_t\|P)$, $t \geq 1$, to obtain the claim. □

Proof of Corollary 10. By the triangle inequality,

$$M_{ts} := \|F_t - G_s\|_\infty - \|F - G\|_\infty \leq \|G_s - G\|_\infty + \|F_t - F\|_\infty\,,$$

so that $\mathbb{E}M_{ts} \leq \sqrt{\pi/2t} + \sqrt{\pi/2s}$, by the same argument as in the proof of Corollary 9. Furthermore, the map

$$(x_1, \dots, x_t, y_1, \dots, y_s) \mapsto \sup_{x \in \mathbb{R}} \left| \frac{1}{t} \sum_{i=1}^{t} I(x_i \leq x) - \frac{1}{s} \sum_{i=1}^{s} I(y_i \leq x) \right|,$$

satisfies the bounded differences property with parameters $L_1 = \cdots = L_t = 1/t$ and $L_{t+1} = \cdots = L_{t+s} = 1/s$. Therefore, McDiarmid's inequality implies

$$\mathbb{P}\big(|M_{ts} - \mathbb{E}M_{ts}| \geq u\big) \leq 2\exp(-2tsu^2/(s+t)), \quad u > 0.$$

It now follows similarly as in the proof of Corollary 9 that M_{ts} is sub-Gaussian with parameter $2(t+s)/ts$. Applying Corollary 6 leads to the bound $\mathbb{P}(\exists t, s \geq 1 : M_{ts} \geq \gamma_{ts}) \leq \delta/2$. Furthermore, from Corollary 8, $\mathbb{P}(\exists t, s \geq 1 : M_{ts} \leq -\kappa_{ts}) \leq \delta/2$. Applying a union bound leads to the claim.

To prove the validity of the test, notice simply that under the null $F = G$, the aforementioned bound reduces to $\mathbb{P}(\exists t, s \geq 1 : \|F_t - G_s\|_\infty \geq \gamma_{ts}) \leq \delta/2$. □

4.D.2 Proofs from Subsection 4.4.2

Proof of Corollary 11 The proof of Theorem 7 (equation (16)) of Gretton et al. (2012) yields the expectation bound

$$\mathbb{E}\big|\mathrm{MMD}(P_t, Q_s) - \mathrm{MMD}(P, Q)\big| \leq 2\Big[(B/t)^{1/2} + (B/s)^{1/2}\Big].$$

Further, the following concentration bound follows from equation (15) of Gretton et al. (2012):

$$\mathbb{P}\big(\big|\mathrm{MMD}(P_t, Q_s) - \mathbb{E}[\mathrm{MMD}(P, Q)]\big| \geq u\big) \leq 2\exp\left\{-\frac{tsu^2}{2B(t+s)}\right\}, \quad u > 0.$$

Therefore, Lemma 32 implies that $\mathrm{MMD}(P_t, Q_s)$ is $\frac{8B(t+s)}{ts}$-sub-Gaussian. Finally, MMD is an IPM by its definition, and is therefore convex by Lemma 31. Combining these facts with Corollary 6, applied to the process $M_{ts} = \mathrm{MMD}(P_t, Q_s) - \mathrm{MMD}(P, Q)$, leads to the bound

$$\mathbb{P}(\exists t, s \geq 1 : \mathrm{MMD}(P_t, Q_s) - \mathrm{MMD}(P, Q) \geq \gamma_{ts}) \leq \delta/2.$$

To obtain an upper confidence sequence, notice that the set $\mathcal{J} = \{f \in \mathcal{H} : \|f\|_{\mathcal{H}} \leq 1\}$ consists of functions taking values in the interval $[-\sqrt{B}, \sqrt{B}]$. Indeed, for all $f \in \mathcal{J}$, $x \in \mathbb{R}^d$,

$$|f(x)| = |\langle f, K(x, \cdot)\rangle_{\mathcal{H}}| \leq \|f\|_{\mathcal{H}} \|K(x, \cdot)\|_{\mathcal{H}} \leq \sqrt{K(x,x)} \leq \sqrt{B}.$$

Applying Corollary 8, we thus have

$$\mathbb{P}(\exists t, s \geq 1 : \mathrm{MMD}(P_t, Q_s) - \mathrm{MMD}(P, Q) \leq -2\sqrt{B}\kappa_{ts}) \leq \delta/2.$$

Applying a union bound leads to the claim. □

Thus far we have studied the plugin estimator $\mathrm{MMD}^2(P_t, Q_s)$, which is known to be a biased estimator of $\mathrm{MMD}^2(P, Q)$. The following unbiased estimator, which we state only in the case $t = s$, is widely-used and is obtained by replacing the V-statistic in (4.31) by the following U-Statistic

$$\widehat{M_t^2} = \frac{1}{t(t-1)} \sum_{i \neq j} \widetilde{J}(Z_i, Z_j), \tag{4.44}$$

where $\widetilde{J}((x,y),(x',y')) = K(x,x') + K(y,y') - K(x,y') - K(x',y)$. The process $\widehat{M_t}$ does not admit a simple characterization as a convex functional of the empirical measures P_t and Q_t, thus Theorem 15 and Proposition 21 cannot be directly applied. U-Statistics are, however, known to be reverse martingales, as discussed in Section 4.4.2, implying that $\widehat{M_t^2}$ is a reverse martingale. While this does not imply that $\widehat{M_t} - \mathrm{MMD}(P, Q)$ is a reverse submartingale, Theorem 15 can be applied directly to the mean-zero process $\widehat{M_t^2} - \mathrm{MMD}^2(P, Q)$.

Proposition 22. Under the same conditions as Corollary 11, we have for all $\delta \in (0, 1)$,

$$\mathbb{P}\left(\exists t \geq 1 : \widehat{M_t^2} \geq \mathrm{MMD}^2(P, Q) + 16B\sqrt{\frac{1}{t-1}\Big[\log \ell(\log_2 t) + \log(1/\delta)\Big]}\right) \leq \delta.$$

Unlike equation (4.32), Proposition 22 does not lead to a confidence sequence for $\mathrm{MMD}^2(P, Q)$ scaling at the rate $O(\log\log t/t)$ when $P = Q$. We therefore recommend the use of the plugin estimator $\mathrm{MMD}(P_t, Q_s)$ and Corollary 11 when a confidence sequence is needed in practice.

Proof of Proposition 22. By Theorem 10 of Gretton et al. (2012) and Lemma 32, one can infer similarly as in the proof of Corollary 11 that $\widehat{M_t^2}$ is $(32B^2/\underline{t})$-sub-Gaussian, where $\underline{t} = \lfloor t/2 \rfloor$. The claim then follows by the same proof technique as Theorem 15 and Corollary 6, using the fact that $\widehat{M_t^2}$ is a reverse martingale, and using the inequality $\lceil \lfloor t/2 \rfloor / 2 \rceil \geq (t-1)/4$. □

We close this section with a proof of Proposition 18.

Proof of Proposition 18. Let $\nu \in \mathcal{P}(\mathcal{X})$ be any fixed reference measure. By Mercer's Theorem (see for instance Christmann and Steinwart (2008), Theorem 4.49), a continuous, symmetric, and positive definite kernel h admits the representation

$$h(x,y) = \sum_{i=0}^{\infty} \lambda_i \psi_i(x)\psi_i(y), \quad x,y \in \mathcal{X},$$

where $(\lambda_i)_{i\geq 0} \subseteq \ell^2 \equiv \ell^2(\mathbb{N}_0)$ is the sequence of eigenvalues corresponding to the Hilbert-Schmidt operator $f \in L^2(\nu) \mapsto \int h(\cdot,y)f(y)d\nu(y)$, and $(\psi_i)_{i\geq 0} \subseteq L^2(\nu)$ is a corresponding sequence of eigenfunctions. It follows that one may write for any $\mu \in \mathcal{P}(\mathcal{X})$,

$$\sqrt{\Phi(\mu)} = \sqrt{\iint h(x,y)d\mu(x)d\mu(y)} = \sqrt{\sum_{i=0}^{\infty} \lambda_i \left(\int \psi_i d\mu\right)^2} = \left\| \left(\sqrt{\lambda_i} \int \psi_i d\mu\right)_{i\geq 0} \right\|_{\ell^2},$$

The right-hand side of the above display is a composition of the convex map $\|\cdot\|_{\ell^2}$ with the affine map $\mu \in V \mapsto \left(\sqrt{\lambda_i} \int \psi_i d\mu\right)_{i\geq 0} \in \ell^2$, where V is the vector space of finite signed measures on $(\mathcal{X}, \mathbb{B}(\mathcal{X}))$. It follows that the functional $\sqrt{\Phi(\cdot)}$ is itself convex. Since this functional is also nonnegative, and the square function is convex and increasing on $\mathbb{R}_+$, it is straightforward to verify that the functional $\Phi(\cdot)$ is likewise convex. The claim then follows from Theorem 14. □

4.D.3 Proofs from Subsection 4.4.3

Proof of Corollary 12. We will make use of the Kantorovich duality (cf. Section 4.2.1) to show that the map

$$\begin{aligned} G : (x_1, \dots, x_t, y_1, \dots, y_s) &\in \mathcal{X}^{t+s} \\ &\mapsto \mathcal{T}_c\left(\frac{1}{t}\sum_{i=1}^t \delta_{x_i}, \frac{1}{s}\sum_{j=1}^s \delta_{y_j}\right) = \sup_{(f,g)\in\mathcal{M}_c} \frac{1}{t}\sum_{i=1}^t f(x_i) + \frac{1}{s}\sum_{j=1}^s g(y_j), \end{aligned}$$

satisfies the bounded differences property. This generalizes the one-sample analogue proven for instance by Weed and Bach (2019) (Proposition 20). Let $1 \leq k \leq t$, and $x_1, \widetilde{x}_1, \dots, x_t, \widetilde{x}_t \in \mathcal{X}$ be such that $\widetilde{x}_i = x_i$ for all $i \neq k$. Furthermore, let $y_1, \dots, y_s \in \mathcal{X}$. Let $(f_0, g_0) \in \mathcal{M}_c$ denote optimal Kantorovich potentials satisfying

$$\mathcal{T}_c\left(\frac{1}{t}\sum_{i=1}^t \delta_{x_i}, \frac{1}{s}\sum_{j=1}^s \delta_{y_j}\right) = \frac{1}{t}\sum_{i=1}^t f_0(x_i) + \frac{1}{s}\sum_{j=1}^s g_0(y_j).$$

Furthermore, recall from Section 4.2.1 that we may (and do) choose (f_0, g_0) such that $0 \leq f_0 \leq \Delta$ and $-\Delta \leq g_0 \leq 0$. Then,

$$\begin{aligned} &G(x_1, \dots, x_t, y_1, \dots, y_s) - G(\widetilde{x}_1, \dots, \widetilde{x}_t, y_1, \dots, y_s) \\ &\quad\leq \frac{1}{t}\sum_{i=1}^t f_0(x_i) + \frac{1}{s}\sum_{j=1}^s g_0(y_j) - \frac{1}{t}\sum_{i=1}^t f_0(\widetilde{x}_i) - \frac{1}{s}\sum_{j=1}^s g_0(y_j) \end{aligned}$$

$$\leq \frac{1}{t}[f_0(x_k) - f_0(\widetilde{x}_k)] \leq \Delta/t.$$

Repeating a symmetric argument, we obtain

$$|G(x_1, \dots, x_t, y_1, \dots, y_s) - G(\widetilde{x}_1, \dots, \widetilde{x}_t, y_1, \dots, y_s)| \leq \Delta/t.$$

We similarly have that for all $\widetilde{y}_1, \dots, \widetilde{y}_s \in \mathcal{X}$, satisfying $\widetilde{y}_i = y_i$ for all $i \neq k$,

$$|G(x_1, \dots, x_t, y_1, \dots, y_s) - G(x_1, \dots, x_t, \widetilde{y}_1, \dots, \widetilde{y}_s)| \leq \Delta/s.$$

We deduce that G satisfies the bounded differences property with parameters $L_1 = \dots = L_t = \Delta/t$ and $L_{t+1} = \dots L_{t+s} = \Delta/s$. McDiarmid's inequality then implies

$$\mathbb{P}\big(|\mathcal{T}_c(P_t, Q_s) - \mathbb{E}\mathcal{T}_c(P_t, Q_s)| \geq u\big) \leq 2\exp\left\{-\frac{2tsu^2}{(t+s)\Delta^2}\right\}, \quad u > 0.$$

It follows that $\mathcal{T}_c(P_t, Q_s)$ is $\frac{2\Delta^2(t+s)}{ts}$-sub-Gaussian by Lemma 32. Since $\mathcal{T}_c$ is convex, applying Corollary 6 yields

$$\mathbb{P}\Bigg(\exists t, s \geq 1 : \mathcal{T}_c(P_t, Q_s) - \mathcal{T}_c(P, Q) \geq \alpha_{c,ts} + 2\Delta\sqrt{\frac{2(t+s)}{ts}\Big[\log g(\log_2 t + \log_2 s) + \log(2/\delta)\Big]}\Bigg) \leq \delta/2.$$

Furthermore, Corollary 8 and the Kantorovich duality immediately lead to the bound

$$\mathbb{P}(\exists t, s \geq 1 : \mathcal{T}_c(P_t, Q_s) - \mathcal{T}_c(P, Q) \leq -\Delta\kappa_{ts}) \leq \delta/2.$$

The claim follows. □

4.D.4 Proofs from Subsection 4.4.4

Proof of Proposition 19. We repeat a similar stitching argument as that of the proof of Theorem 15. Let $N_t = \mathrm{KL}(P_t \| P)$, $t \geq 1$. (N_t) forms a reverse submartingale by Lemma 31 and Theorem 14, implying by Jensen's inequality that for any integer $t_1 \geq 1$, $\big(\exp(t_1\lambda_{t_1}N_t)\big)_{t=1}^\infty$ is a reverse submartingale. Therefore, following along similar lines as the proof of Theorem 15, and applying Theorem 13, we have for all $y > 0$ and all integers $t_0 \geq 1$,

$$\begin{aligned}\mathbb{P}\Big(\exists t \geq t_0 : N_t \geq y\Big) &= \mathbb{P}\Big(\exists t \geq t_0 : \exp(\lambda_{t_1}t_1N_t) \geq \exp(\lambda_{t_1}t_1y)\Big) \\ &\leq \mathbb{E}[\exp(-yt_1\lambda_{t_1} + t_1\lambda_{t_1}N_{t_0})] \leq \exp(-yt_1\lambda_{t_1})G_{k,t_0}(t_1\lambda_{t_1}/t_0).\end{aligned}$$

Now, letting $u_j = 2^j$ for all integers $j \geq 0$, and $\gamma_t = \frac{1}{\lambda_t t}\log\big(\delta^{-1}G_{k,\lfloor t/2\rfloor}(2\lambda_t)\ell(\log_2 t)\big)$, we obtain

$$\mathbb{P}\left(\exists t \geq 1 : N_t \geq \gamma_t\right) \leq \mathbb{P}\left(\bigcup_{j=0}^\infty \Big\{\exists t \in \{u_j, \dots, u_{j+1}\} : N_t \geq \gamma_t\Big\}\right)$$

$$\begin{aligned}
&\leq \mathbb{P}\left(\bigcup_{j=0}^{\infty}\left\{\exists t \in \{u_j,\dots,u_{j+1}\} : N_t \geq \gamma_{u_{j+1}}\right\}\right)\\
&\leq \sum_{j=0}^{\infty} \exp\left\{-u_{j+1}\gamma_{u_{j+1}}\lambda_{u_{j+1}}\right\} G_{k,u_j}(2\lambda_{u_{j+1}}) \leq \sum_{j=0}^{\infty}\frac{\delta}{\ell(j+1)} \leq \delta,
\end{aligned}$$

where on the final line, we used the fact that $G_{k,t}(\lambda)$ increases with t for all fixed $k \geq 2, \lambda \in [0,1]$ (Guo and Richardson (2020), Lemma 1). The claim follows. □

Proof of Corollary 13. By Berend and Kontorovich (2013), Eq. (17), and references therein, we have

$$\mathbb{P}\Big(\, \|P_t - P\|_{\mathrm{TV}} - \mathbb{E}\big[\, \|P_t - P\|_{\mathrm{TV}}\,\big] \geq u\Big) \leq 2\exp(-2tu^2), \quad u > 0,$$

implying that $\|P_t - P\|_{\mathrm{TV}}$ is $(2/t)$-sub-Gaussian. Furthermore, Lemma 7 of Kamath et al. (2015) implies

$$\mathbb{E}[\|P_t - P\|_{\mathrm{TV}}] \leq \sqrt{\frac{k}{2\pi t}} + 2\left(\frac{k}{t}\right)^{\frac{3}{4}}.$$

Finally, $\|\cdot\|_{\mathrm{TV}}$ forms a convex divergence by Lemma 31. The claim now follows by Corollary 6. □

4.D.5 Proofs from Subsection 4.4.5

We shall make use of the following result due to Bobkov and Götze (1999).

Lemma 34 (Bobkov and Götze (1999), Theorem 1.3)**.** *Let d denote a metric on $\mathcal{X}$. Then, a measure $\mu \in \mathcal{P}_1(\mathcal{X})$ satisfies the $T_1(\sigma^2)$ inequality with respect to d if and only if $f(X)$ is σ^2-sub-Gaussian for all functions $f : \mathcal{X} \to \mathbb{R}$ which are 1-Lipschitz with respect to d.*

Proof of Proposition 20. To prove part (i), it is straightforward to show that the map

$$G : (x_1,\dots,x_t) \in \mathbb{R}^{t\times d} \longmapsto \left\|\frac{1}{t}\sum_{i=1}^{t}(\delta_{x_i} - P)\right\|_{\mathrm{TV}}^{\sigma},$$

satisfies the bounded differences property. Indeed, given $1 \leq j \leq t$, let $x_1, \widetilde{x}_1, \dots, x_t, \widetilde{x}_t \in \mathbb{R}^d$, such that $x_i = \widetilde{x}_i$ for all $i \neq j$. Then, the triangle inequality implies

$$\begin{aligned}
|G(x_1,\dots,x_t) - G(\widetilde{x}_1,\dots,\widetilde{x}_t)| &\leq \sup_{A\in\mathbb{B}(\mathbb{R}^d)}\left|\left(\frac{1}{t}\sum_{i=1}^{t}\delta_{x_i}\star\mathcal{K}_\sigma\right)(A) - \left(\frac{1}{t}\sum_{i=1}^{t}\delta_{\widetilde{x}_i}\star\mathcal{K}_\sigma\right)(A)\right|\\
&\leq \frac{1}{t}\sup_{A\in\mathbb{B}(\mathbb{R}^d)}\int_A |K_\sigma(x - x_j) - K_\sigma(x - \widetilde{x}_j)|\,dx\\
&\leq \frac{1}{t}\sup_{A\in\mathbb{B}(\mathbb{R}^d)}\int_A \Big[K_\sigma(x - x_j) + K_\sigma(x - \widetilde{x}_j)\Big]dx \leq 2/t.
\end{aligned}$$

Therefore, by McDiarmid's Inequality (Theorem 16), we have

$$\mathbb{P}\Big(\big| \|P_t - P\|_{\mathrm{TV}}^\sigma - \mathbb{E}\, \|P_t - P\|_{\mathrm{TV}}^\sigma \big| \geq u \Big) \leq 2\exp(-tu^2/2), \quad u > 0.$$

It follows from Lemma 32 that $\|P_t - P\|_{\mathrm{TV}}^\sigma$ is $8t$-sub-Gaussian. Furthermore, Goldfeld et al. (2020) show that $\mathbb{E}\, \|P_t - P\|_{\mathrm{TV}}^\sigma \leq c_d t^{-1/2}/\sqrt{2}$. Finally, the Total Variation distance is convex by Lemma 31, thus $\|\cdot\|_{\mathrm{TV}}^\sigma$ is also convex. The first claim now follows from Corollary 6.

To prove the second claim, we show similarly as Niles-Weed and Rigollet (2022) that the map

$$G : (x_1, \dots, x_t) \in \mathbb{R}^{t \times d} \longmapsto tW_1^\sigma \left(\frac{1}{t} \sum_{i=1}^t \delta_{x_i}, P \right),$$

is Lipschitz with respect to the metric $c_t(x, y) := \sum_{i=1}^t \|x_i - y_i\|_2$ on $\mathbb{R}^{d \times t}$, where $x = (x_1, \dots, x_t), y = (y_1, \dots, y_t) \in \mathbb{R}^{d \times t}$. Let $\mathcal{J}$ denote the set of 1-Lipschitz functions on $\mathbb{R}^d$, and recall that the W_1 distance coincides with the IPM generated by $\mathcal{J}$, by the Kantorovich-Rubinstein duality. We thus have, by the triangle inequality for W_1,

$$\begin{aligned}
|G(x) - G(y)| &\leq tW_1 \left(\left(\frac{1}{t} \sum_{i=1}^t \delta_{x_i} \right) \star \mathcal{K}_\sigma, \left(\frac{1}{t} \sum_{i=1}^t \delta_{y_i} \right) \star \mathcal{K}_\sigma \right) \\
&= t \sup_{f \in \mathcal{J}} \int f d \left(\frac{1}{t} \sum_{i=1}^t (\delta_{x_i} \star \mathcal{K}_\sigma - \delta_{y_i} \star \mathcal{K}_\sigma) \right) \\
&= \sup_{f \in \mathcal{J}} \sum_{i=1}^t \big[(f \star K_\sigma)(x_i) - (f \star K_\sigma)(y_i) \big] \\
&= \sup_{f \in \mathcal{J}} \sum_{i=1}^t \int [f(x_i - z) - f(y_i - z)] K_\sigma(z) dz \\
&\leq \sum_{i=1}^t \|x_i - y_i\|_2 \int K_\sigma(z) dz = c_t(x, y).
\end{aligned}$$

We deduce that G is 1-Lipschitz with respect to c_t. Furthermore, by Gozlan and Léonard (2010), Proposition 1.9, the product measure $P^{\otimes t}$ satisfies the $T_1(t\tau^2)$ inequality over $\mathbb{R}^{d \times t}$ with respect to c_t. Therefore, $G(X_1, \dots, X_t) = tW_1^\sigma(P_t, P)$ is $(t\tau^2)$-sub-Gaussian by Lemma 34, i.e. $W_1^\sigma(P_t, P)$ is (τ^2/t)-sub-Gaussian. Furthermore, $\mathbb{E} W_1^\sigma(P_t, P) \leq C_d t^{-1/2}/\sqrt{2}$ by Goldfeld et al. (2020). Applying Corollary 6 leads to the claim. □

Proof of Corollary 14. Since P is supported in $[-1, 1]^d$, Proposition 5 of Polyanskiy and Wu (2016) implies

$$|h(P_t \star \mathcal{K}_\sigma) - h(P \star \mathcal{K}_\sigma)| \leq \frac{1}{2\sigma^2} \left(|\mu_t - \mu| + 2\sqrt{d} W_1^\sigma(P_t, P) \right), \tag{4.45}$$

where $\mu_t = \int x dP_t(x)$ and $\mu = \int x dP(x)$. Notice that P is 1-sub-Gaussian by Hoeffding's Lemma, and thus also satisfies the $T_1(1)$ inequality (by Lemma 34). By Corollary 16 (see also the discussion thereafter), we have

$$\forall t \geq 1 : |\mu_t - \mu| \leq 2\sqrt{\frac{1}{t}\Big[\log \ell(\log_2 t) + \log(4/\delta)\Big]}, \quad \text{with probability at least } 1-\delta/2,$$

and by Corollary 20,

$$\forall t \geq 1 : W_1^\sigma(P_t, P) \leq \frac{C_d}{\sqrt{t}} + 2\sqrt{\frac{1}{t}\Big[\log \ell(\log_2 t) + \log(2/\delta)\Big]}, \quad \text{with probability at least } 1-\delta/2.$$

By a union bound and equation (4.45), it follows that with probability at least $1-\delta$, we have uniformly in $t \geq 1$,

$$\begin{aligned}
|h(P_t \star \mathcal{K}_\sigma) - h(P \star \mathcal{K}_\sigma)| &\leq \frac{1}{2\sigma^2}\Big\{|\mu_t - \mu| + 2\sqrt{d}W_1^\sigma(P_t, P)\Big\} \\
&\leq \frac{1}{2\sigma^2}\left\{(2+4\sqrt{d})\sqrt{\frac{1}{t}\Big[\log \ell(\log_2 t) + \log(4/\delta)\Big]} + \frac{2\sqrt{d}C_d}{\sqrt{t}}\right\} \\
&\leq \frac{3\sqrt{d}}{\sigma^2}\sqrt{\frac{1}{t}\Big[\log \ell(\log_2 t) + \log(4/\delta)\Big]} + \frac{\sqrt{d}C_d}{\sqrt{t}\sigma^2},
\end{aligned}$$

as claimed. □

4.D.6 Proofs from Subsection 4.4.6

Proof of Corollary 15. Notice that

$$\sup_{f\in\mathcal{F}} |R(f) - R_t(f)| = \sup_{f\in\mathcal{F}}\left|\int I(f(x) \neq y) d(P-P_t)(x,y)\right| = D_{\mathcal{J}}(P_t\|P),$$

where $D_{\mathcal{J}}$ is the IPM generated by the class $\mathcal{J} = \{(x,y) \mapsto I(f(x) \neq y) : f \in \mathcal{F}\}$. Since the functions in $\mathcal{J}$ are uniformly bounded by 1, if follows by the same argument as in the proof of Corollary 9 that that $D_{\mathcal{J}}(P_t\|P)$ is $(2/t)$-sub-Gaussian. Furthermore, a standard symmetrization argument (see for instance equation (4.18) of Wainwright (2019)) implies

$$\mathbb{E}[D_{\mathcal{J}}(P_t\|P)] \leq 2\mathcal{R}_t(\mathcal{J}) = \mathcal{R}_t(\mathcal{F}),$$

where the final equality follows from Lemma 3.4 of Mohri, Rostamizadeh, and Talwalkar (2018). By Corollary 6, we deduce

$$\mathbb{P}\left(\exists t \geq 1 : D_{\mathcal{J}}(P_t\|P) \geq \mathcal{R}_{\bar{t}}(\mathcal{F}) + 2\sqrt{\frac{2}{t}\Big[\log \ell(\log_2 t) + \log(1/\delta)\Big]}\right) \leq \delta,$$

which readily implies the first claim. To prove the second, abbreviate $\widehat{\mathcal{R}}_t(\mathcal{F})$ by $\widehat{\mathcal{R}}_t$, and let

$$\widehat{\mathcal{R}}_t^i = \mathbb{E}_{\boldsymbol{\epsilon}_{t+1}}\left[\sup_{f\in\mathcal{F}} \frac{1}{t}\left|\sum_{\substack{j=1 \\ j\neq i}}^{t+1} \epsilon_j f(X_j)\right|\right], \quad i = 1, \dots, t+1.$$

Then,

$$\begin{aligned}\widehat{\mathcal{R}}_{t+1} &= \mathbb{E}_{\boldsymbol{\epsilon}_{t+1}}\left[\sup_{f\in\mathcal{F}} \frac{1}{t+1}\left|\sum_{j=1}^{t+1} \epsilon_j f(X_j)\right|\right] = \mathbb{E}_{\boldsymbol{\epsilon}_{t+1}}\left[\sup_{f\in\mathcal{F}} \frac{1}{t+1}\left|\sum_{i=1}^{t+1} \frac{1}{t}\sum_{\substack{j=1\\ j\neq i}}^{t+1} \epsilon_j f(X_j)\right|\right] \\ &\le \frac{1}{t+1}\sum_{i=1}^{t+1} \mathbb{E}_{\boldsymbol{\epsilon}_{t+1}}\left[\sup_{f\in\mathcal{F}} \frac{1}{t}\left|\sum_{\substack{j=1\\ j\neq i}}^{t+1} \epsilon_j f(X_j)\right|\right] = \frac{1}{t+1}\sum_{i=1}^{t+1} \widehat{\mathcal{R}}_t^i,\end{aligned}$$

implying that $(\widehat{\mathcal{R}}_t)$ satisfies the leave-one-out property in equation (4.16). It follows from Proposition 14 that $(\widehat{\mathcal{R}}_t)$ is a reverse submartingale with respect to the exchangeable filtration $(\mathcal{E}_t^X)$. Thus, Theorem 15 and Corollary 6 can be invoked with (N_t) replaced by $(\widehat{\mathcal{R}}_t)$. Furthermore, it can again be deduced as before that $\widehat{\mathcal{R}}_t$ is $(8/t)$-sub-Gaussian, and has mean $\mathcal{R}_t$. Corollary 6 thus leads to the claim. □

4.D.7 Proofs from Subsection 4.4.7

Let $\mathcal{A}_\gamma$ be a γ-cover of $\mathbb{S}_\star^{d-1}$ of size N_γ. By a straightforward covering argument, notice that for any $\nu \in \mathbb{S}_\star^{d-1}$, there exists $\nu_0 \in \mathcal{A}_\gamma$ such that $\|\nu-\nu_0\|_* \le \gamma$ thus

$$\nu^\top X = (\nu-\nu_0)^\top X + \nu_0^\top X \le \gamma\,\|X\| + \nu_0^\top X,$$

whence

$$\|X\| = \sup_{\nu\in\mathbb{S}_\star^{d-1}} \nu^\top X \le \gamma\,\|X\| + \max_{\nu\in\mathcal{A}_\gamma} \nu^\top X,$$

implying that $\|X\| \le \frac{1}{1-\gamma}\max_{\nu\in\mathcal{A}_\gamma} \nu^\top X$. We deduce that for all $\lambda \ge 0$,

$$\begin{aligned}\mathbb{E}\left[\exp\left(\lambda\,\|\mu_t-\mu\|\right)\right] &\le \mathbb{E}\left[\exp\left(\frac{\lambda}{1-\gamma}\max_{\nu\in\mathcal{A}_\gamma}\nu^\top(\mu_t-\mu)\right)\right] \\ &\le \sum_{\nu\in\mathcal{A}_\gamma} \mathbb{E}\left[\exp\left(\frac{\lambda}{1-\gamma}\nu^\top(\mu_t-\mu)\right)\right] \\ &= \sum_{\nu\in\mathcal{A}_\gamma}\left(\mathbb{E}\left[\exp\left(\frac{\lambda}{t(1-\gamma)}\nu^\top(X-\mu)\right)\right]\right)^t \\ &\le N_\gamma \exp\left(t\psi\left(\frac{\lambda}{t(1-\gamma)}\right)\right).\end{aligned}$$

where we extend the definition of the convex function ψ to $\mathbb{R}_+$ by setting $\psi(\lambda)=\infty$ for all $\lambda \ge \lambda_{\max}$. We deduce that an upper bound on the cumulant generating function of $\|\mu_t-\mu\|$ is given by

$$\bar{\psi}_t(\lambda) = \log N_\gamma + t\psi\left(\frac{\lambda}{t(1-\gamma)}\right), \quad \lambda \ge 0,$$

It is readily seen that for all $x \in \mathbb{R}$ and all $\lambda \geq 0$,

$$\bar{\psi}_t^*(x) = -\log N_\gamma + t\psi^*\big((1-\gamma)x\big), \quad \big(\bar{\psi}_t^*\big)^{-1}(\lambda) = \frac{1}{1-\gamma}(\psi^*)^{-1}\left(\frac{\lambda + \log N_\gamma}{t}\right).$$

Finally, notice that the functional $\Phi(Q) = \big\|\int x dQ(x) - \mu\big\|$ is convex, thus we may apply Theorem 15 to deduce that for all $\delta \in (0,1)$,

$$\mathbb{P}\left\{\exists t \geq 1 : \|\mu_t - \mu\| \geq \frac{1}{1-\gamma}(\bar{\psi}^*)^{-1}\left(\frac{\log \ell(\log_2 t) + \log(1/\delta) + \log N_\gamma}{\lceil t/2 \rceil}\right)\right\} \leq \delta.$$

The claim follows. □

4.D.8 Proofs from Subsection 4.4.8

Proof of Corollary 17. Let $\eta, \alpha > 1$, and set $u_k = \lceil \eta^k \rceil$ for all $k \geq 0$. Define $\ell(k) = (1 \vee k^\alpha)\zeta(\alpha)$, where $\zeta(\alpha) = \sum_{k=1}^\infty \frac{1}{k^\alpha}$ and $\alpha > 1$. From the proof of Theorem 17, for the process $N_t = D(P_t\|P)$ in the special case of sub-Gaussian tails $\psi_t(\lambda) = \lambda\mathbb{E}(N_t) + \lambda^2\sigma^2/2t$, it can be seen that $\mathbb{P}(A_k) \leq \delta/\ell(k+1)$, where, for all $k = 0, 1, \ldots$,

$$A_k = \left\{\exists u_k \leq t \leq u_{k+1} : N_t > \mathbb{E}\left(N_{\lceil t/\lceil \eta \rceil \rceil}\right) + \sqrt{\frac{2\sigma^2}{\lceil t/\lceil \eta \rceil \rceil}\Big[\log \ell(\log_\eta t) + \log(1/\delta)\Big]}\right\}.$$

Thus, by definition of ℓ and by the first Borel-Cantelli Lemma, we have $\mathbb{P}(\limsup_{k\to\infty} A_k) = 0$. Therefore,

$$\mathbb{P}\left\{N_t \leq \mathbb{E}\left(N_{\lceil t/\lceil \eta \rceil \rceil}\right) + \sqrt{\frac{2\sigma^2}{\lceil t/\lceil \eta \rceil \rceil}\Big[\log \ell(\log_\eta t) + \log(1/\delta)\Big]} \text{ eventually}\right\} = 1.$$

Note that for all $t \geq \eta$,

$$\log \ell(\log_\eta t) = \alpha \log\log_\eta t + \log\zeta(\alpha) = \alpha\log\log t - \alpha\log\log\eta + \log\zeta(\alpha).$$

Therefore, we have almost surely,

$$\limsup_{t\to\infty} \frac{D(P_t\|P)}{\mathbb{E}\left(N_{\lceil t/\lceil \eta \rceil \rceil}\right) + \sqrt{\frac{2\sigma^2}{\lceil t/\lceil \eta \rceil \rceil}\Big[\alpha\log\log t - \alpha\log\log\eta + \log\zeta(\alpha) + \log(1/\delta)\Big]}} \leq 1,$$

whence, by assumption on $\mathbb{E}N_t$ and by the fact that δ, α, η are fixed, we obtain

$$\limsup_{t\to\infty} \frac{D(P_t\|P)}{\sqrt{\frac{2\sigma^2}{\lceil t/\lceil \eta \rceil \rceil}\alpha\log\log t}} \leq 1 \quad \text{a.s.}$$

Since we may choose η and α arbitrarily close to 1, the claim follows. □

Part II

Smooth Optimal Transport

Chapter 5

Plugin Estimation of Smooth Optimal Transport Maps

5.1 Introduction

In this chapter, we momentarily pause our study of optimal transport costs, and turn our attention to the related question of estimating Brenier maps. As we have already discussed in Chapter 1, Brenier maps have found a wide range of recent methodological applications, in which they need to be estimated from data. Our aim will be to define several natural estimators of Brenier maps, and to show that they are minimax optimal.

Let us briefly recall the definition of a Brenier map. Given two absolutely continuous probability distributions $P, Q \in \mathcal{P}_{\mathrm{ac}}(\Omega)$, with support contained in a compact set $\Omega \subseteq \mathbb{R}^d$, the Brenier map, or (quadratic) optimal transport map, from P to Q is the P-almost everywhere uniquely defined solution T_0 to the Monge problem,

$$\underset{T \in \mathcal{T}(P,Q)}{\operatorname{argmin}} \int_\Omega \|x - T(x)\|^2 \, dP(x), \tag{5.1}$$

where we recall that $\mathcal{T}(P, Q)$ is the set of transport maps between P and Q. As we have seen in Theorem 2, the unique solution T_0 to the above optimization problem takes the form of the gradient of a convex function φ_0, i.e. $T_0 = \nabla\varphi_0$. Unlike T_0, the map φ_0 is not unique, and any allowable choice of such convex function is referred to as a *Brenier potential*.

For the statistical applications we have in mind, the distributions P and Q are typically unknown, and we assume the practitioner has access to i.i.d. samples $X_1, \dots, X_n \sim P$ and $Y_1, \dots, Y_m \sim Q$. Our aim is to derive estimators $\widehat{T}_{nm}$ which achieve the minimax rate of convergence[1], under the loss function

$$\big\|\widehat{T}_{nm} - T_0\big\|_{L^2(P)}^2 = \int_\Omega \big\|\widehat{T}_{nm}(x) - T_0(x)\big\|^2 dP(x). \tag{5.2}$$

[1] Here and throughout, minimax rate-optimality is tacitly understood up to polylogarithmic factors.

The theoretical study of such estimators was recently initiated by Hütter and Rigollet (2021), who proved that for any estimator $\widehat{T}_{nm}$ with $n = m$,

$$\sup_{(P,Q)} \mathbb{E}\big\|\widehat{T}_{nm} - T_0\big\|_{L^2(P)}^2 \gtrsim n^{-\frac{2\alpha}{2(\alpha-1)+d}} \vee \frac{1}{n}, \tag{5.3}$$

where the supremum is taken over all pairs of distributions (P, Q) admitting densities bounded away from zero over a compact set Ω, for which T_0 lies in an α-Hölder ball for some $\alpha \geq 1$, and satisfies a key curvature condition **A1($\boldsymbol{\lambda}$)** which we define below. The lower bound (5.3) is reminiscent of, but generally faster than, the classical $n^{-2\alpha/(2\alpha+d)}$ minimax rate of estimating an α-Hölder continuous nonparametric regression function (Tsybakov, 2008), and is shown by Hütter and Rigollet (2021) to be achievable up to a polylogarithmic factor. Nevertheless, their estimator is computationally intractable in general dimension, and their work leaves open the question of developing practical optimal transport map estimators which achieve comparable risk.

In this chapter, we establish the minimax optimality of several natural and intuitive estimators of optimal transport maps, several of which have already been proposed in the statistical optimal transport literature, but have resisted sharp statistical analyses thus far. We focus on the following two classes of plugin estimators.

(i) **Empirical Estimators.** When no smoothness assumptions are placed on P and Q, it is natural to study the plugin estimator based on the empirical measures

$$P_n = \frac{1}{n}\sum_{i=1}^n \delta_{X_i}, \quad \text{and} \quad Q_m = \frac{1}{m}\sum_{j=1}^m \delta_{Y_j}.$$

In the special case $n = m$, there is an optimal transport map T_{nm} from P_n to Q_m, and more generally there is an optimal coupling of these measures. While the in-sample estimator T_{nm} is only defined over the support of P_n, we readily obtain estimators defined over the entire domain by casting the extension problem as one of nonparametric regression. We show how linear smoothers and least-squares estimators can be used to interpolate T_{nm}, leading to estimators $\widehat{T}_{nm}$ defined over Ω. Such estimators are new in the literature to the best of our knowledge, and achieve the minimax rate for estimating Lipschitz optimal transport maps T_0.

(ii) **Smooth Estimators.** In order to obtain faster rates of convergence when P and Q admit smooth densities p and q, we next analyze the risk of the unique optimal transport map between kernel or wavelet density estimators of p and q. In contrast to our empirical optimal transport map estimators, we show that such smooth plugin estimators are able to take advantage of additional regularity of the densities p and q, and achieve minimax-optimal rates when these densities are Hölder smooth.

While our emphasis is on optimal transport maps, it will turn out that our analysis also leads to new convergence rates for various plugin estimators of the 2-Wasserstein distance, which could not have been deduced from our results in Chapter 3. Some of the results developed in this chapter will also play a very important role in our development of central limit theorems for the 2-Wasserstein distance, a topic which we take up in Chapter 7 below.

Our Contributions. The primary contributions of this chapter are summarized as follows.

(i) In Sections 5.2 and 5.3, we develop new stability bounds which relate the risk of plugin transport map estimators to the plugin density estimation risk, as measured in the Wasserstein distance. These stability bounds are quite general and enable the analysis of flexible, practical transport map estimators. The risk of density estimation under the Wasserstein distance has been extensively studied (Niles-Weed and Berthet, 2022; Divol, 2021), and our stability bounds enable us to leverage this past work. Additionally, our stability bounds enable the analysis of plugin estimators of the Wasserstein distance, once again relating the risk in this problem to the plugin density estimation risk.

(ii) We build upon our stability bounds to analyze the risk of empirical, kernel-based and wavelet-based transport map estimators in both the one-sample setup (where the source distribution is known exactly, and the target distribution is sampled) and the two-sample setup (where both the source and target distributions are sampled). The rates we obtain are minimax optimal. For example, suppose that $\widehat{T}_n$ is the optimal transport map from P to $\widehat{Q}_n$, where $\widehat{Q}_n$ is a wavelet-estimator over the domain $[0,1]^d$. Then, whenever P and Q admit $(\alpha-1)$-Hölder densities and satisfy several additional conditions, we show that,

$$\mathbb{E}\big\|\widehat{T}_n - T_0\big\|_{L^2(P)}^2 \lesssim \begin{cases} n^{-\frac{2\alpha}{2(\alpha-1)+d}}, & d \geq 3 \\ (\log n)^2/n, & d = 2 \\ 1/n, & d = 1. \end{cases} \tag{5.4}$$

As we saw in Chapter 1, the Hölder smoothness of T_0 is typically expected to be of one degree greater than that of p and q, and thus our estimator achieves the minimax lower bound (5.3) when these densities are $(\alpha-1)$-Hölder smooth, for any $\alpha > 1^2$. In the two-sample setting, we develop analogous minimax-optimal analyses, for the empirical plugin estimator (Propositions 24–26) as well as for kernel-based and wavelet-based plugin estimators (Theorems 22–20) when P and Q admit Hölder-smooth densities. In the latter case, as we discuss further in the sequel, we avoid complications that arise in the optimal transport problem due to boundary effects by working over the d-dimensional flat torus.

(iii) In each of the above settings, we complement our results with upper bounds on the risk of plugin estimators of the Wasserstein distance. For instance, in the smooth setting discussed above, we show that,

$$\mathbb{E}\big|W_2^2(P,\widehat{Q}_n) - W_2^2(P,Q)\big| \lesssim \left(\frac{1}{n}\right)^{\frac{2\alpha}{2(\alpha-1)+d}} \vee \frac{1}{\sqrt{n}}. \tag{5.5}$$

[2] As discussed in Appendix E of Hütter and Rigollet (2021), the minimax lower bound (5.3) also holds under such smoothness conditions on the densities p and q, as opposed to smoothness conditions on T_0.

> We also develop analogous results in the one and two-sample settings, for various empirical and smooth plugin estimators. The above display implies that the Wasserstein distance can be estimated at the parametric rate in any dimension, provided that the smoothness exponent α of the underlying map satisfies $2(\alpha+1) > d$. In this regime, it will turn out that $W_2^2(P, \widehat{Q}_n)$ also enjoys a $\sqrt{n}$-central limit theorem centered around its population counterpart; see Theorem 26 of Chapter 7 below.

Related Work. The two recent works of Hütter and Rigollet (2021) and Gunsilius and Xu (2021) establish $L^2(P)$ convergence rates for transport map estimators. Gunsilius and Xu (2021) derives upper bounds on the risk of a plugin estimator for Brenier potentials, obtained via kernel density estimation of p and q. This analysis results in suboptimal convergence rates for the optimal transport map T_0 itself. We show in this work that such plugin estimators do in fact achieve the optimal convergence rate when the sampling domain is the d-dimensional torus.

Building upon a construction of del Barrio et al. (2020), a consistent estimator of T_0 was obtained by De Lara, González-Sanz, and Loubes (2021) under mild assumptions, by regularizing a piecewise constant approximation of the empirical optimal transport map T_n. We do not know if quantitative convergence rates can be obtained for their estimator under stronger assumptions. Beyond these works, a wide range of heuristic estimators have been proposed in the literature (Perrot et al., 2016; Nath and Jawanpuria, 2020; Makkuva et al., 2020), but their theoretical properties remain unknown to the best of our knowledge.

Rates of convergence for the problem of estimating Wasserstein distances have arguably received more attention than that of estimating optimal transport maps; see Chapter 3 and references therein. Although we showed in that chapter that the empirical plugin estimator of the Wasserstein distance is minimax optimal up to polylogarithmic factor under no assumptions on P and Q, it becomes suboptimal when P and Q have smooth densities. Niles-Weed and Berthet (2022) derive the minimax rate of estimating smooth densities under the Wasserstein distance, and we build upon their results, together with those of Divol (2021), to characterize the risk of our density plugin estimators (cf. Sections 5.2.3, 5.3.3, and 5.3.4).

Finally, let us emphasize that during the final stages of preparation of a preprint containing the main results of this chapter, we became aware of the independent work of Deb and Sen (2021), and of the most recently revised version of the work of Ghosal and Sen (2022). These papers bound the risk of certain plugin optimal transport map estimators that are closely related to those in our work. In particular, assuming for simplicity that $n = m$, they show that an estimator derived from the empirical plugin optimal transport coupling achieves the $n^{-\left(\frac{1}{2}\wedge\frac{2}{d}\right)}$ convergence rate under the squared $L^2(P_n)$ loss up to polylogarithmic factors. Our work establishes an analogous result using a distinct proof, but further shows that empirical estimators achieve this rate in squared $L^2(P)$ norm, once suitably extended using nonparametric smoothers. We also sharpen this result to the rate $n^{-\left(1\wedge\frac{2}{d}\right)}$ under additional conditions. Deb and Sen (2021) also analyze the convergence rate of plugin estimators based on wavelet and kernel density estimation. Their work shows that such estimators can achieve, for instance, the faster rate $n^{-\left(\frac{1}{2}\vee\frac{\alpha}{d+2(\alpha-1)}\right)}$, when the underlying densities lie in a $(\alpha-1)$-Hölder ball for some $\alpha > 1$. While this upper bound illustrates an improvement over empirical estimators in the presence of smoothness, it scales at a quadratically slower rate than the minimax rate (5.3). In contrast, our work shows that wavelet density plugin estimators do in fact achieve the minimax rate $n^{-\left(1\vee\frac{2\alpha}{d+2(\alpha-1)}\right)}$ (up to a polylogarithmic factor when $d = 2$). We also extend this result to kernel density estimators, using a significantly different proof strategy than Deb and Sen (2021). Finally, we emphasize that our sharp analysis of estimators for the Wasserstein distance allows us to deduce that their bias is of lower order than their variance when $2(\alpha+1) > d$, which is a key component in our derivation of their limiting distribution in Chapter 7 below.

5.2 The One-Sample Problem

Throughout this section, we let $P \in \mathcal{P}_{\mathrm{ac}}(\Omega)$ denote a known distribution, and $Q \in \mathcal{P}_{\mathrm{ac}}(\Omega)$ denote an unknown distribution from which an i.i.d. sample $Y_1, \dots, Y_n \sim Q$ is observed. Let p and q denote their respective densities, and let $T_0 = \nabla\varphi_0$ denote the unique optimal transport map from P to Q, with respect to a convex Brenier potential φ_0. We also denote by $\phi_0 = \|\cdot\|^2 - 2\varphi_0$ and $\psi_0 = \|\cdot\|^2 - 2\varphi_0^*$ the Kantorovich potentials induced by φ_0. We assume here and throughout the remainder of this chapter, except where otherwise specified, that the set $\Omega \subseteq \mathbb{R}^d$ satisfying the following condition.

(S1) Ω is a compact, convex set with nonempty interior such that $\Omega \subseteq [0,1]^d$.

Notice that once Ω is assumed compact, the final assumption in condition **(S1)** can always be guaranteed by rescaling. Furthermore, since $\mathrm{diam}(\Omega) \leq d$ under condition **(S1)**, we may assume without loss of generality that $-d \leq \phi_0 \leq 0$ and $0 \leq \psi_0 \leq d$ over Ω (Villani (2003), Remark 1.13).

Unlike the two-sample case which we discuss in Section 5.3, there exist canonical estimators of T_0 when the source distribution P is known. Indeed, since P is absolutely continuous, Brenier's Theorem implies that there exists a unique optimal transport map $\widehat{T}$ between P and any estimator $\widehat{Q}$ of Q, and we analyze two such examples below. We first take $\widehat{Q}$ to be the empirical measure of Q in Section 5.2.2, and show that the resulting estimator $\widehat{T}$ achieves the minimax risk of estimating Lipschitz optimal transport maps, under essentially no smoothness conditions on the underlying measures. In Section 5.2.3, we then take $\widehat{Q}$ to be a density estimator, leading to an estimator $\widehat{T}$ achieving faster rates of convergence when Q admits a smooth density. In both cases, our analysis will hinge upon known upper bounds on the risk of $\widehat{Q}$ under the Wasserstein distance, by invoking a key stability bound which we turn to first.

5.2.1 A General Stability Bound

The main technical result of this section will be stated under the following curvature condition.

A1(λ) The Brenier potential φ_0 is a convex function such that $\varphi_0 \in \mathcal{C}^2(\Omega)$ and $(1/\lambda)I_d \preceq$

$\nabla^2\varphi_0(x) \preceq \lambda I_d$ for all $x \in \Omega$.

It can be seen that whenever condition **A1(λ)** holds for φ_0, the same bounds also hold for φ_0^*. Therefore, condition **A1(λ)** implies that T_0 is λ-bi-Lipschitz over Ω. Furthermore, we note the following simple result of Gigli (2011).

Lemma 35. *Assume $\varphi_0 \in \mathcal{C}^2(\Omega)$ and $\gamma^{-1} \le p, q \le \gamma$ for some $\gamma > 0$. Then, there exists a constant $\lambda > 0$, depending only on γ and $\|\varphi_0\|_{\mathcal{C}^2(\Omega)}$, such that φ_0 is $(1/\lambda)$-strongly convex.*

Lemma 35 implies that when P and Q both admit densities which are bounded from above and below over Ω, the second inequality of condition **A1(λ)** implies the first, up to suitably inflating λ. Under this condition, we prove the following stability bounds in Appendix 5.B.

Theorem 18. Let $P, Q \in \mathcal{P}_{\mathrm{ac}}(\Omega)$, and assume condition **A1(λ)** holds for some $\lambda > 0$. For any $\widehat{Q} \in \mathcal{P}(\Omega)$, let $\widehat{T} = \nabla\widehat{\varphi}$ be the unique optimal transport map from P to $\widehat{Q}$. Then,

$$\frac{1}{\lambda}\|\widehat{T} - T_0\|_{L^2(P)}^2 \le W_2^2(P, \widehat{Q}) - W_2^2(P, Q) - \int \psi_0 d(\widehat{Q} - Q) \le \lambda W_2^2(\widehat{Q}, Q). \tag{5.6}$$

We make several remarks regarding Theorem 18.

- Caffarelli's regularity theory (cf. Theorem 3) provides sufficient conditions on the smoothness of P, Q and $\partial\Omega$ for assumption **A1(λ)** to hold, albeit for a non-universal constant $\lambda > 0$. We note, however, that our assumption is considerably weaker. For instance, condition **A1(λ)** is satisfied whenever P and Q differ by a location transformation, irrespective of the regularity or positivity of their Lebesgue densities.
- We show in Section 7.2 that, under weaker assumptions than those of Theorem 18, the map $\psi_0 - \mathbb{E}_Q[\psi_0(Y)]$ is the efficient influence function of the functional $Q \in \mathcal{P}(\Omega) \mapsto W_2^2(P, Q)$ with respect to the tangent space $L_0^2(Q)$. It follows that the linear functional

  $$L(\widehat{Q}) = \int \psi_0 d(\widehat{Q} - Q) \tag{5.7}$$

  is the first-order term in the von Mises expansion of $W_2^2(P, \widehat{Q})$ around $W_2^2(P, Q)$. The upper bound of Theorem 18 implies that the remainder of this expansion decays quadratically in the topology induced by W_2, a fact which we shall use to derive upper bounds and limit laws for plugin estimators of the Wasserstein distance. This fact combined with the lower bound of Theorem 18 further implies the following remarkable equivalence,

  $$\frac{1}{\lambda}\|\widehat{T} - T_0\|_{L^2(P)} \le W_2(\widehat{Q}, Q) \le \|\widehat{T} - T_0\|_{L^2(P)}. \tag{5.8}$$

  Notice that the second inequality always holds due to the fact that $(\widehat{T}, T_0)_{\#}P$ is a coupling of $\widehat{Q}$ and Q. Equation (5.8) thus shows that the transport cost of this coupling is within a universal factor of being optimal, when the curvature condition **A1(λ)** is in force. We use this result to obtain upper bounds on the risk of one-sample plugin estimators $\widehat{T}$ by appealing to the corresponding risk of $\widehat{Q}$ under the Wasserstein distance.

- When $d = 1$, it is easy to see by direct calculation that the inequalities (5.8) hold with equality, with $\lambda = 1$, even without assumption **A1(λ)**. For multivariate measures, weaker analogues of equation (5.8), in which the left-hand side admits an exponent greater than unity, have previously been derived by Mérigot, Delalande, and Chazal (2020); Delalande and Merigot (2023). Those works adopted a weaker assumption than ours, however.
- Suppose that, in addition to the assumptions of Theorem 18, the measures Q and $\widehat{Q}$ are both absolutely continuous with respect to the Lebesgue measure, with respective densities q and $\widehat{q}$ which satisfy $\gamma^{-1} \leq q, \widehat{q} \leq \gamma$ over Ω, for some $\gamma > 0$. In this setting, it was shown by Peyre (2018) that the 2-Wasserstein distance is equivalent to the negative-order homogeneous Sobolev norm $\|\cdot\|_{\dot{H}^{-1}(\Omega)}$, in the sense that, under suitable conditions on Ω,
$$\gamma^{-1}\|\widehat{q} - q\|^2_{\dot{H}^{-1}(\Omega)} \lesssim W_2^2(\widehat{Q}, Q) \lesssim \gamma\|\widehat{q} - q\|^2_{\dot{H}^{-1}(\Omega)}. \tag{5.9}$$
Theorem 18 and the above display then imply
$$\frac{1}{\lambda\gamma}\|\widehat{\varphi} - \varphi_0\|^2_{\dot{H}^1(\Omega)} \lesssim W_2^2(P, \widehat{Q}) - W_2^2(P, Q) - \int \psi_0 d(\widehat{Q} - Q) \lesssim \lambda\gamma\|\widehat{q} - q\|^2_{\dot{H}^{-1}(\Omega)}.$$
It follows from the upper bound that $W_2^2(P, \cdot)$, when viewed as a functional of $\widehat{q}$, is Fréchet differentiable at q in the $\dot{H}^{-1}(\Omega)$ topology. It moreover implies that this functional is strongly convex and smooth with respect to the duality of the spaces $\dot{H}^{-1}(\Omega)$ and $\dot{H}^1(\Omega)$.
- Theorem 18 is stated in a form which is sufficient for our purposes, however it is not the most general result possible. On the one hand, the assumption of boundedness on Ω is superfluous: Theorem 18 continues to hold if Ω is an unbounded, closed, and convex set, such as the entire Euclidean space $\mathbb{R}^d$. It follows, for instance, that Theorem 18 is applicable whenever P and Q are strongly log-concave measures, in which case assumption **A1(λ)** holds by Caffarelli's contraction theorem (Caffarelli, 2000). On the other hand, assumption **A1(λ)** can be weakened in the following way: the first inequality of display (5.6) holds under the mere condition $\nabla^2\varphi_0 \preceq \lambda I_d$, whereas the second holds when $\nabla^2\varphi_0 \succeq \lambda^{-1} I_d$.
- Finally, one may also infer from Theorem 18 and the Kantorovich duality that,
$$\frac{1}{2\lambda}\|\nabla\widehat{\varphi} - \nabla\varphi_0\|^2_{L^2(P)} \leq \int(\varphi_0 - \widehat{\varphi})dP + \int(\varphi_0^* - \widehat{\varphi}^*)d\widehat{Q} \leq \frac{\lambda}{2}\|\nabla\widehat{\varphi} - \nabla\varphi_0\|^2_{L^2(P)}. \tag{5.10}$$
Equation (5.10) is a direct analogue of a stability bound proven by Hütter and Rigollet (2021, Proposition 10), who show that similar inequalities hold when the measure $\widehat{Q}$ appearing in the above display is replaced by Q. Their result assumes, however, that $\widehat{\varphi}$ itself satisfies condition **A1(λ)**. In contrast, we do not place any conditions on the estimator $\widehat{T}$ beyond it being the optimal transport map from P to $\widehat{Q}$. This will permit our study of transport map estimators which are potentially nonsmooth but easy to compute, as we show next.

5.2.2 Upper Bounds for One-Sample Empirical Estimators

Recall that $Q_n = (1/n)\sum_{i=1}^n \delta_{Y_i}$ denotes the empirical measure. Since P is known and absolutely continuous, a natural estimator for T_0 is the optimal transport map T_n from P to Q_n, defined by

$$T_n = \operatorname*{argmin}_{T\in\mathcal{T}(P,Q_n)} \int \|x - T(x)\|^2 \, dP(x). \tag{5.11}$$

By Brenier's Theorem, the minimizer T_n in the above display exists and is uniquely determined P-almost everywhere. The optimization problem (5.11) is sometimes known as the semi-discrete optimal transport problem, for which efficient numerical solvers are well-studied (Mérigot, 2011; Levy and Schwindt, 2018).

In view of the stability bound in Theorem 18, the risk of T_n may be related to that of the empirical measure Q_n under the Wasserstein distance. For instance, from the work of Fournier and Guillin (2015) we obtain the following bound, under no assumptions beyond **(S1)**,

$$\mathbb{E}W_2^2(Q_n, Q) \lesssim \kappa_n := \begin{cases} n^{-1/2}, & d \leq 3 \\ n^{-1/2}\log n, & d = 4 \\ n^{-2/d}, & d \geq 5. \end{cases} \tag{5.12}$$

The following bound on the risk of T_n is now an immediate consequence of Theorem 18, together with the fact that the functional L in equation (5.7) satisfies $\mathbb{E}[L(Q_n)] = 0$.

Corollary 18. *Let $P, Q \in \mathcal{P}_{\mathrm{ac}}(\Omega)$ and assume condition **A1(λ)** holds. Then,*

$$\mathbb{E}\big\|T_n - T_0\big\|_{L^2(P)}^2 \asymp_\lambda \mathbb{E}\big[W_2^2(P, Q_n) - W_2^2(P, Q)\big] \asymp_\lambda \mathbb{E}W_2^2(Q_n, Q) \lesssim \kappa_n.$$

When $d \geq 5$, Corollary 18 implies that the empirical estimator T_n achieves the minimax lower bound (5.3) for estimating Lipschitz transport maps T_0. On the other hand, when $1 \leq d \leq 4$, the rate κ_n does not improve beyond $n^{-1/2}$, unlike the minimax lower bound (5.3) of Hütter and Rigollet (2021), which scales as fast as $1/n$. This observation does not imply that the plugin estimator T_n is minimax suboptimal, since equation (5.3) holds under stronger assumptions than those of Corollary 18. In particular, it assumes that these distributions admit densities which are bounded away from zero, and thus have connected support. In contrast, Corollary 18 applies to measures P and Q with possibly disconnected support, for which our upper bound of κ_n cannot generally be improved up to a logarithmic factor—similar considerations are discussed for the convergence rate of the empirical measure by Bobkov and Ledoux (2019) when $d = 1$, and more generally by Niles-Weed and Berthet (2022).

Nevertheless, when we further assume that Q has a positive density, the result of Corollary 18 can be strengthened to match the minimax rate of Hütter and Rigollet (2021) even for $d \leq 4$. For instance, it is well-known (cf. Ajtai, Komlós, and Tusnády (1984), Ledoux (2019)) that, when Q is the uniform distribution on $[0,1]^d$, Q_n achieves the following faster rate,

$$\mathbb{E}W_2^2(Q_n,Q) \lesssim \begin{cases} n^{-1}, & d=1 \\ n^{-1}\log n, & d=2 \\ n^{-2/d}, & d\geq 3. \end{cases} \tag{5.13}$$

Such a result is also known to hold for any measure Q admitting positive density over a compact subset of the real line (Bobkov and Ledoux, 2019), or over the flat torus (Divol, 2021). Inspired by the latter result and by the work of Niles-Weed and Berthet (2022), we prove an analogue of equation (5.13) for arbitrary measures supported on the unit hypercube, at the expense of an inflated polylogarithmic factor when $d=2$.

Corollary 19. *Let $P,Q \in \mathcal{P}_{\mathrm{ac}}([0,1]^d)$ and assume that condition **A1(λ)** holds. Assume further that $\gamma^{-1} \leq q \leq \gamma$ over $[0,1]^d$, for some $\gamma > 0$. Then,*

$$\mathbb{E}\big\|T_n - T_0\big\|_{L^2(P)}^2 \asymp \mathbb{E}\big[W_2^2(P,Q_n) - W_2^2(P,Q)\big] \asymp \mathbb{E}W_2^2(Q_n,Q) \lesssim \bar{\kappa}_n := \begin{cases} n^{-1}, & d=1 \\ \frac{(\log n)^2}{n}, & d=2 \\ n^{-2/d}, & d\geq 3. \end{cases}$$

Under the assumptions of Corollary 19, we deduce that the plugin estimator T_n is minimax optimal for all $d \geq 1$, up to a polylogarithmic factor when $d=2$. The scale of this factor is further discussed following the statement of Theorem 20.

This result also provides a sharper bound on the bias of $W_2^2(P,Q_n)$ than could have been deduced from Chapter 3. Furthermore, Corollaries 18–19 can be extended to recover our risk bounds from Chapter 3 under stronger conditions, though with an improved rate of convergence when P approaches Q in Wasserstein distance.

Corollary 20. *Let $P,Q \in \mathcal{P}_{\mathrm{ac}}(\Omega)$, and assume condition **A1(λ)** holds. Then,*

$$\mathbb{E}\big|W_2^2(P,Q_n) - W_2^2(P,Q)\big| \lesssim_\lambda \mathbb{E}W_2^2(Q_n,Q) + n^{-\frac{1}{2}} \lesssim \kappa_n. \tag{5.14}$$

If we further assume that $\Omega = [0,1]^d$ and $\gamma^{-1} \leq q \leq \gamma$ over Ω for some $\gamma > 0$, then

$$\mathbb{E}\big|W_2^2(P,Q_n) - W_2^2(P,Q)\big| \lesssim_{\lambda,\gamma} \bar{\kappa}_n + W_2(P,Q)n^{-\frac{1}{2}}. \tag{5.15}$$

Equation (5.15) exhibits an upper bound on the risk of $W_2^2(P,Q_n)$ that interpolates between the fast rate $\bar{\kappa}_n$ when $W_2(P,Q) \lesssim n^{-1/2}$, and the rate κ_n which is optimal when the distance between P and Q is unconstrained (cf. Theorem 12 of Chapter 3). We defer the proofs of Corollaries 19–20 to Appendix 5.C.

5.2.3 Upper Bounds for One-Sample Wavelet Estimators

While the empirical estimator in the previous section achieves the minimax rate of estimating Lipschitz optimal transport maps, we do not generally expect it to achieve faster rates of convergence if T_0 is assumed to enjoy further regularity. We instead show that such improvements can be achieved when Q admits a smooth density q, and when the empirical measure Q_n is replaced by the distribution $\widehat{Q}_n$ of a density estimator $\widehat{q}_n$. Specifically, define

$$\widehat{T}_n = \operatorname*{argmin}_{T \in \mathcal{T}(P, \widehat{Q}_n)} \int \|x - T(x)\|^2 \, dP(x). \tag{5.16}$$

We focus on the case where $\widehat{q}_n$ is a wavelet density estimator, for which sharp risk estimates under the Wasserstein distance have been established by Niles-Weed and Berthet (2022). In order to appeal to their results, we assume that the sampling domain is the unit hypercube $\Omega = [0,1]^d$. In Section 5.3.4, we also extend some of the results of this section to the case where Ω is a generic domain with smooth boundary.

We briefly introduce notation from the theory of wavelets, and refer the reader to Appendix A for a detailed summary and references. To define a basis over the unit cube Ω, we focus on the boundary-corrected N-th Daubechies wavelet system, for an integer $N \geq 2$, as introduced by Cohen, Daubechies, and Vial (1993). In short, given an integer $j_0 \geq \log_2 N$, their construction leads to respective families of scaling and wavelet functions

$$\Phi^{\mathrm{bc}} = \{\zeta^{\mathrm{bc}}_{j_0 k} : 0 \leq k \leq 2^{j_0} - 1\}, \quad \Psi^{\mathrm{bc}}_j = \{\xi^{\mathrm{bc}}_{jk\ell} : 0 \leq k \leq 2^{j_0} - 1, \ell \in \{0,1\}^d \backslash \{0\}\}, \quad j \geq j_0,$$

such that $\Psi^{\mathrm{bc}} = \Phi^{\mathrm{bc}} \cup \bigcup_{j=j_0}^{\infty} \Psi^{\mathrm{bc}}_j$ forms an orthonormal basis of $L^2(\Omega)$, with the property that Φ^{bc} spans all polynomials of degree at most $N-1$ over Ω. Given a probability distribution $Q \in \mathcal{P}_{\mathrm{ac}}(\Omega)$ admitting density $q \in L^2(\Omega)$, one then has

$$q = \sum_{\xi \in \Psi^{\mathrm{bc}}} \beta_\xi \xi = \sum_{\zeta \in \Phi^{\mathrm{bc}}} \beta_\zeta \zeta + \sum_{j=j_0}^{\infty} \sum_{\xi \in \Psi^{\mathrm{bc}}_j} \beta_\xi \xi, \quad \text{where} \quad \beta_\xi = \int \xi dQ, \ \xi \in \Psi^{\mathrm{bc}},$$

where the series converges at least in $L^2(\Omega)$. The standard truncated wavelet estimator of q (Kerkyacharian and Picard, 1992) with a truncation level $J_n \geq j_0 > 0$ is then given by

$$\widetilde{q}^{(\mathrm{bc})}_n = \sum_{\xi \in \Psi^{\mathrm{bc}}} \widehat{\beta}_\xi \xi = \sum_{\zeta \in \Phi^{\mathrm{bc}}} \widehat{\beta}_\zeta \zeta + \sum_{j=j_0}^{J_n} \sum_{\xi \in \Psi^{\mathrm{bc}}_j} \widehat{\beta}_\xi \xi, \quad \text{where} \quad \widehat{\beta}_\xi = \int \xi dQ_n, \ \xi \in \Psi^{\mathrm{bc}}.$$

Notice that $\widetilde{q}^{(\mathrm{bc})}_n$ is permitted to take on negative values, in which case it does not define a probability density. We instead define the final density estimator $\widehat{q}_n \equiv \widehat{q}^{(\mathrm{bc})}_n$ by

$$\widehat{q}^{(\mathrm{bc})}_n = \frac{\widetilde{q}^{(\mathrm{bc})}_n I(\widetilde{q}^{(\mathrm{bc})}_n \geq 0)}{\int \widetilde{q}^{(\mathrm{bc})}_n I(\widetilde{q}^{(\mathrm{bc})}_n \geq 0)}, \quad \text{over } \Omega, \tag{5.17}$$

and we denote by $\widehat{Q}^{(\mathrm{bc})}_n$ the distribution induced by $\widehat{q}^{(\mathrm{bc})}_n$. We drop all superscripts "bc" in the sequel whenever the choice of wavelet system is unambiguous. Niles-Weed and Berthet (2022) bounded the Wasserstein risk of a wavelet density estimator obtained from a distinct modification of $\widetilde{q}_n$. By appealing to L^∞ concentration inequalities for wavelet density estimators (Masry, 1997), we show in Appendix B.1.0.4 that their result carries over to the estimator

$\widehat{q}_n$. Equipped with this result, we arrive at the following bound on the risk of the estimator $\widehat{T}_n \equiv \widehat{T}_n^{(\mathrm{bc})}$ defined in equation (5.16), and of the corresponding plugin estimator of the squared Wasserstein distance. Here and throughout, we will adopt the notation

$$\mathcal{C}^\alpha(\Omega; M) := \left\{ f \in \mathcal{C}^\alpha(\Omega) : \|f\|_{\mathcal{C}^\alpha(\Omega)} \leq M \right\}, \tag{5.18}$$

$$\mathcal{C}^\alpha(\Omega; M, \gamma) := \left\{ f \in \mathcal{C}^\alpha(\Omega) : \|f\|_{\mathcal{C}^\alpha(\Omega)} \leq M, f \geq 1/\gamma \text{ over } \Omega \right\}, \tag{5.19}$$

for any $M, \gamma > 0$.

Theorem 19 (One-Sample Wavelet Estimators)**.** Let $\alpha > 1$ and $M, \gamma > 0$. Let $P, Q \in \mathcal{P}_{\mathrm{ac}}([0,1]^d)$, and assume the density q satisfies $q \in \mathcal{C}^{\alpha-1}([0,1]^d; M, \gamma)$. Let $2^{J_n} \asymp n^{1/(d+2(\alpha-1))}$. Then, the following assertions hold.

(i) (Optimal Transport Maps) Assume φ_0 satisfies condition **A1(λ)** for some $\lambda > 0$. Then, there exists a constant $C > 0$ depending on $M, \lambda, \gamma, \alpha$ such that,

$$\mathbb{E}\big\|\widehat{T}_n - T_0\big\|_{L^2(P)}^2 \leq C R_{T,n}(\alpha), \quad \text{where } R_{T,n}(\alpha) := \begin{cases} 1/n, & d = 1 \\ (\log n)^2/n, & d = 2 \\ n^{-\frac{2\alpha}{2(\alpha-1)+d}}, & d \geq 3. \end{cases}$$

(ii) (Wasserstein Distances) Assume that for some $\lambda > 0$, $\varphi_0^* \in \mathcal{C}^{\alpha+1}([0,1]^d; \lambda)$. Then, there exists a constant $C > 0$ depending on $M, \lambda, \gamma, \alpha$ such that,

$$\big|\mathbb{E}W_2^2(P, \widehat{Q}_n) - W_2^2(P, Q)\big| \leq C R_{T,n}(\alpha),$$

$$\mathbb{E}\big|W_2^2(P, \widehat{Q}_n) - W_2^2(P, Q)\big|^2 \leq \left[C R_{T,n}(\alpha) + \sqrt{\frac{\mathrm{Var}_Q[\psi_0(Y)]}{n}} \right]^2.$$

Theorem 19 requires smoothness assumptions on both the density q and the potential φ_0^*; in particular, the assumption of Theorem 19(ii) requires both $q \in \mathcal{C}^{\alpha-1}(\Omega)$ and $\varphi_0^* \in \mathcal{C}^{\alpha+1}(\Omega)$. Caffarelli's regularity theory (Theorem 3) suggests that the former condition on q should be sufficient to imply the latter condition on φ_0^*, but such results cannot be invoked here due to the lack of smoothness of the boundary of the unit cube $[0,1]^d$. Even if the above analysis could be adapted to a domain Ω with smooth boundary, the lack of uniformity in Caffarelli's global regularity theory would prevent the bounds in Theorem 19 from holding uniformly in P and Q, in the absence of a smoothness condition on φ_0^*. We refer to Appendix E of Hütter and Rigollet (2021) for related discussions. In Proposition 29 of Appendix 5.G, we will show that an analogue of Theorem 19 holds merely under smoothness conditions on p and q when Ω is the d-dimensional torus $\mathbb{T}^d$, which enjoys the global regularity result of Theorem 4. Here, we instead impose smoothness conditions on both φ_0^* and q, in which case $\widehat{T}_n$ achieves the minimax rate (5.3) of estimating an α-Hölder optimal transport map.

Theorem 19(ii) also proves that the bias of $W_2^2(P, \widehat{Q}_n)$ achieves the same convergence rate, as does its risk when $d \geq 2(\alpha + 1)$. In the high-smoothness regime $d < 2(\alpha + 1)$, the risk of $W_2^2(P, \widehat{Q}_n)$, in squared loss, does not generally improve beyond the parametric rate $1/n$, except when $\mathrm{Var}_Q[\psi_0(Y)]$ vanishes. Using Lemma 37 in Appendix 5.A, the latter quantity is bounded above by $W_2^2(P, Q)$ up to a constant, so Theorem 19(ii) also implies

$$\mathbb{E}\big|W_2^2(P, \widehat{Q}_n) - W_2^2(P, Q)\big| \lesssim_{M,\gamma,\lambda,\alpha} R_{T,n}(\alpha) + \frac{W_2(P, Q)}{\sqrt{n}}. \tag{5.20}$$

We briefly highlight the main components of the proof of Theorem 19. Both assertions are proven by combining the stability results of Theorem 18 with the bound $\mathbb{E}W_2^2(\widehat{Q}_n, Q) \lesssim R_{T,n}(\alpha)$, which is stated formally in Lemma 106, and extends a result due to Niles-Weed and Berthet (2022). In particular, Theorem 19(i) follows immediately from the equivalence (5.8). Our proof of Theorem 19(ii) additionally requires us to analyze the evaluation $L(\widehat{Q}_n)$ of the linear functional L defined in equation (5.7), for which we prove the following.

Lemma 36. *Assume the same conditions as Theorem 19(ii). Then,*

$$\mathbb{E}[L(\widehat{Q}_n)] = O\left(2^{-2J_n\alpha}\right), \quad \mathrm{Var}\left[L(\widehat{Q}_n)\right] = \frac{1}{n}\mathrm{Var}_Q[\psi_0(Y)] + O\left(\frac{2^{-2J_n\alpha}}{n}\right),$$

where the implicit constants depend only on $M, \gamma, \lambda, \alpha$.

Lemma 36 shows that the bias of $L(\widehat{Q}_n)$ scales quadratically faster than the traditional bias of $\widehat{Q}_n$ in estimating an $(\alpha - 1)$-Hölder density, which is known to be of order $2^{-J_n(\alpha-1)}$. We obtain the faster rate $2^{-2J_n\alpha}$ due to the assumed $(\alpha+1)$-Hölder smoothness of the potential φ_0^*. The proofs of Theorem 19 and Lemma 36 are deferred to Appendix 5.D.

Remark 2 (Adaptive Estimation)**.** When constructing the estimator $\widehat{T}_n$, we assumed that the smoothness parameter α is known, and used it to tune the truncation parameter J_n. It is also possible to construct an adaptive estimator, however. Niles-Weed and Berthet (2022, Theorem 2) derived an adaptive density estimator $\widehat{Q}_n^\circ$ which achieves the minimax rate of estimating Q under the Wasserstein distance, up to polylogarithmic factors. It is then natural to define a plugin estimator of T_0 as the unique optimal transport map from P to $\widehat{Q}_n^\circ$. By reasoning similarly as in the proof of Theorem 19(i), this estimator has an $L^2(P)$ risk of order $R_{T,n}(\alpha)$, up to polylogarithmic factors, and does not require knowledge of α.

5.3 The Two-Sample Problem

In this section, we turn to analyzing two-sample estimators when both measures $P, Q \in \mathcal{P}_{\mathrm{ac}}(\Omega)$ are unknown. As in the one-sample case, we study two classes of plugin estimators. The first consists of estimators which interpolate the empirical in-sample optimal transport coupling using nonparametric smoothers. Such estimators will achieve the optimal rate of estimating T_0 when it is Lipschitz. The second class will consist of plugin estimators based on density estimates of P and Q, and will achieve faster rates of convergence when P and Q have smooth densities. As before, our proofs will rely on stability bounds for the two-sample problem, to which we turn our attention first.

5.3.1 Two-Sample Stability Bounds

The stability bounds of Theorem 18 admit the following one-sided extension when both measures P and Q are unknown.

Proposition 23. Let $P, Q \in \mathcal{P}_{\mathrm{ac}}(\Omega)$, and assume condition **A1(λ)** holds for some $\lambda > 0$. Then, for any measures $\widehat{P}, \widehat{Q} \in \mathcal{P}(\Omega)$,

$$\begin{aligned} 0 \leq W_2^2(\widehat{P}, \widehat{Q}) - W_2^2(P, Q) - \int \phi_0 d(\widehat{P} - P) - \int \psi_0 d(\widehat{Q} - Q) \\ \leq \lambda \left[W_2(\widehat{P}, P) + W_2(\widehat{Q}, Q) \right]^2 . \end{aligned} \tag{5.21}$$

The proof is deferred to Appendix 5.E.1. Similarly to Theorem 18, this result shows that the remainder of a first-order expansion of $W_2^2(\widehat{P}, \widehat{Q})$ around $W_2^2(P, Q)$ decays quadratically in the W_2 topology. Unlike Theorem 18, however, we do not generally expect that the lower bound in Proposition 23 can be replaced by a squared distance between (P, Q) and $(\widehat{P}, \widehat{Q})$: for instance, the lower bound of zero is achieved in equation (5.21) when $\widehat{P} = \widehat{Q} \neq Q = P$, even though $\widehat{P}$ may be arbitrarily far from P in Wasserstein distance. This example shows more generally that the bivariate functional $W_2^2(\cdot, \cdot)$ is not strictly convex over $\mathcal{P}_{\mathrm{ac}}(\Omega) \times \mathcal{P}_{\mathrm{ac}}(\Omega)$, unlike the univariate functional $W_2^2(P, \cdot)$ for a fixed absolutely continuous measure P (cf. Theorem 18 and Proposition 7.19 of Santambrogio (2015)).

These observations do not preclude the possibility of replacing the lower bound in Proposition 23 by $\lambda^{-1}\|\widehat{T} - T_0\|_{L^2(P)}^2$, for $\widehat{T}$ the optimal transport map between $\widehat{P}$ and $\widehat{Q}$. We were not able to derive such a result under the stated assumptions, except when these estimators are taken to be empirical measures. We describe this special case next, and show how it may be used to derive estimators of Lipschitz optimal transport maps T_0.

5.3.2 Upper Bounds for Two-Sample Empirical Estimators

Let $X_1, \dots, X_n \sim P$ and $Y_1, \dots, Y_m \sim Q$ denote i.i.d. samples, and define the empirical measures $P_n = (1/n)\sum_{i=1}^n \delta_{X_i}$ and $Q_m = (1/m)\sum_{j=1}^m \delta_{Y_j}$. Though the Monge problem between P_n and Q_m can be infeasible when $n \neq m$, the Kantorovich problem is always feasible, and takes the following form

$$\widehat{\pi} \in \operatorname*{argmin}_{\pi \in \mathcal{Q}_{nm}} \sum_{i=1}^n \sum_{j=1}^m \pi_{ij} \|X_i - Y_j\|^2 ,$$

where $\mathcal{Q}_{nm}$ denotes the set of doubly stochastic matrices $\pi = (\pi_{ij} : 1 \leq i \leq n,\ 1 \leq j \leq m)$, satisfying $\pi_{ij} \geq 0$, $\sum_{i=1}^n \pi_{ij} = 1/m$ and $\sum_{j=1}^m \pi_{ij} = 1/n$. We shall formulate the main stability bound of this section in terms of the quantity

$$\Delta_{nm} = \sum_{i=1}^n \sum_{j=1}^m \widehat{\pi}_{ij} \|T_0(X_i) - Y_j\|^2 .$$

Recall that (κ_n) and $(\overline{\kappa}_n)$ denote the sequences defined in equation (5.12) and Corollary 19 respectively. We obtain the following result, which we prove in Appendix 5.E.3.

Proposition 24. Let $P, Q \in \mathcal{P}_{\mathrm{ac}}(\Omega)$, and assume **A1(λ)** holds for some $\lambda > 0$. Then,

$$\mathbb{E}[\Delta_{nm}] \asymp_\lambda \mathbb{E}\Big[W_2^2(P_n, Q_m) - W_2^2(P, Q)\Big] \lesssim \kappa_{n\wedge m}.$$

If, in addition, $\Omega = [0,1]^d$ and there exists $\gamma > 0$ such that $\gamma^{-1} \leq p, q \leq \gamma$ over Ω, then,

$$\mathbb{E}[\Delta_{nm}] \asymp_\lambda \mathbb{E}\Big[W_2^2(P_n, Q_m) - W_2^2(P, Q)\Big] \lesssim_\gamma \overline{\kappa}_{n\wedge m}.$$

To gain intuition about Proposition 24, it is fruitful to consider the special case $n = m$. In this setting, there exists an optimal transport map T_n from P_n to Q_n, and we may take

$$\widehat{\pi}_{ij} = I(T_n(X_i) = Y_j)/n, \quad \text{for all } 1 \leq i, j \leq n.$$

We then have $\Delta_{nn} = \|T_n - T_0\|^2_{L^2(P_n)}$, and Proposition 24 implies

$$\mathbb{E}\,\|T_n - T_0\|^2_{L^2(P_n)} \asymp \mathbb{E}\Big[W_2^2(P_n, Q_n) - W_2^2(P, Q)\Big]. \tag{5.22}$$

Equation (5.22) is a two-sample analogue of Corollary 18, and shows that the $L^2(P_n)$ risk of the in-sample transport map estimator is of same order as the bias of the two-sample empirical optimal transport cost. While the estimators T_n and $\widehat{\pi}$ are only defined over the support of P_n, we next show how they may be extended to the entire domain Ω. We begin with an estimator inspired by the classical method of nearest-neighbor nonparametric regression (Cover, 1968).

One-Nearest Neighbor Estimator. Define the Voronoi partition generated by $X_1, \dots, X_n$ as

$$V_j = \{x \in \Omega : \|x - X_j\| \leq \|x - X_i\|,\ \forall i \neq j\}, \quad j = 1, \dots, n. \tag{5.23}$$

Then, we define the one-nearest neighbor estimator of T_0 by

$$\widehat{T}^{\mathrm{1NN}}_{nm}(x) = \sum_{i=1}^n \sum_{j=1}^m (n\widehat{\pi}_{ij}) I(x \in V_i) Y_j, \quad x \in \Omega. \tag{5.24}$$

In order to state an upper bound on the convergence rate of $\widehat{T}^{\mathrm{1NN}}_{nm}$, we place the following mild condition on the support Ω. Recall that $\mathcal{L}$ denotes the Lebesgue measure on $\mathbb{R}^d$.

- **(S2)** Ω is a standard set, in the sense that there exist $\epsilon_0, \delta_0 > 0$ such that for all $x \in \Omega$ and $\epsilon \in (0, \epsilon_0)$, we have $\mathcal{L}(B(x,\epsilon) \cap \Omega) \geq \delta_0 \mathcal{L}(B(x, \epsilon))$.

Condition **(S2)** arises frequently in the literature on statistical set estimation (Cuevas and Fraiman, 1997; Cuevas, 2009), and prevents Ω from admitting cusps. Under this condition, we arrive at the following upper bound, which we prove in Appendix 5.F.1.

Proposition 25. Let $P \in \mathcal{P}_{\mathrm{ac}}(\Omega)$ admit a density p such that $\gamma^{-1} \leq p \leq \gamma$ over Ω, for some $\gamma > 0$, and let $Q \in \mathcal{P}_{\mathrm{ac}}(\Omega)$. Assume conditions **A1(λ)** and **(S1)–(S2)** hold. Then,

$$\mathbb{E}\big\|\widehat{T}_{nm}^{\mathrm{1NN}} - T_0\big\|_{L^2(P)}^2 \lesssim_{\lambda,\gamma,\epsilon_0,\delta_0} (\log n)^2 \kappa_{n\wedge m}.$$

Furthermore, if $\Omega = [0,1]^d$ and we additionally assume that $\gamma^{-1} \leq q \leq \gamma$ over Ω, then

$$\mathbb{E}\big\|\widehat{T}_{nm}^{\mathrm{1NN}} - T_0\big\|_{L^2(P)}^2 \lesssim_{\lambda,\gamma,\epsilon_0,\delta_0} (\log n)^2 \overline{\kappa}_{n\wedge m}.$$

Proposition 25 proves that the one-nearest neighbor estimator achieves the minimax rate in equation (5.3), up to a polylogarithmic factor. This result is in stark contrast to standard risk bounds for K-nearest neighbor nonparametric regression, for which the number K of nearest neighbors is typically required to diverge in order to achieve the minimax estimation rate of a Lipschitz continuous regression function (Györfi et al., 2006). Though increasing K reduces the variance of such estimators, in our setting, Propositions 24–25 suggest that the variance of $\widehat{T}_{nm}^{\mathrm{1NN}}$ is already dominated by its large bias, stemming from that of the in-sample coupling $\widehat{\pi}$. Therefore, the choice $K = 1$ is sufficient to obtain a near-optimal rate. While the one-nearest neighbor estimator is simplest to analyze, it is natural to expect that any linear smoother with sufficiently small bandwidth may be used to smooth the in-sample coupling $\widehat{\pi}_{nm}$ and lead to a similar rate.

Convex Least Squares Estimator. Though nearly minimax optimal, the estimator $\widehat{T}_{nm}^{\mathrm{1NN}}$ is typically not the gradient of a convex function, and is therefore not an admissible optimal transport map in its own right. We next show how this property can be enforced using an estimator inspired by nonparametric least squares regression. Let $\mathcal{J}_\lambda$ denote the class of functions $\varphi : \mathbb{R}^d \to \mathbb{R}$ which are convex and have λ-Lipschitz gradients $\nabla\varphi$ over Ω. Define the least squares estimator

$$\widehat{T}_{nm}^{\mathrm{LS}} = \nabla\widehat{\varphi}_{nm}^{\mathrm{LS}}, \quad \text{where } \widehat{\varphi}_{nm}^{\mathrm{LS}} \in \operatorname*{argmin}_{\varphi\in\mathcal{J}_\lambda} \sum_{i=1}^n \sum_{j=1}^m \widehat{\pi}_{ij} \left\|Y_j - \nabla\varphi(X_i)\right\|^2.$$

The computation of the above infinite-dimensional optimization problem can be reduced to that of solving a finite-dimensional quadratic program, by a direct extension of well-known solvers for shape-constrained nonparametric regression with Lipschitz and convex constraints (cf. Seijo and Sen (2011), Mazumder et al. (2019), and references therein). We obtain the following upper bound by a simple extension of Proposition 25.

Proposition 26. Proposition 25 continues to hold when $\widehat{T}_{nm}^{\mathrm{1NN}}$ is replaced by $\widehat{T}_{nm}^{\mathrm{LS}}$.

5.3.3 Upper Bounds for Two-Sample Estimators over $\mathbb{T}^d$

We next study two-sample estimators under stronger smoothness assumptions on P and Q. As discussed in Section 5.3.1, we do not know of a two-sample stability bound for the $L^2(P)$ loss which is analogous to Theorem 18, placing regularity conditions only on the population potential φ_0. Therefore, unlike Theorem 19, in which smoothness conditions on q and φ_0 were sufficient to obtain sharp upper bounds, in the two-sample case our analysis will also rely on the smoothness of *estimators* $\widehat{\varphi}_{nm}$ of the potential φ_0. In order to quantify their regularity, we shall require a uniform analogue of Caffarelli's global regularity theory (Theorem 3(ii)). Since we are unaware of such results for generic compact domains $\Omega \subseteq \mathbb{R}^d$, we instead assume throughout this subsection that Ω is taken to be the d-dimensional torus $\mathbb{T}^d$, thus allowing us to appeal to Theorem 4. We emphasize that our restriction to the torus represents a common approach in the optimal transport literature, whereby numerical methods (Benamou and Brenier, 2000; Loeper and Rapetti, 2005) and theoretical results (Bonnotte, 2013; Guittet, 2003; Santambrogio, 2015) are first derived on the torus before being extended to more generic domains. The torus is an idealized sampling domain, which we believe captures the main qualitative features of our problem, while removing technical issues that arise from boundaries or lack of compactness. As such, it serves as a useful prototype for more general results on compact Euclidean domains with boundaries. In order to illustrate this point, we will prove in the next subsection that our results over the torus extend to generic Euclidean domains Ω, provided that one is willing to assume uniformity in Caffarelli's global regularity theory on Ω.

Though we impose periodicity for technical purposes, we note that optimal transport has recently been used as a methodological tool in several applications involving periodic data, such as high energy physics (cf. Komiske, Metodiev, and Thaler (2019); Komiske et al. (2020), where proton collisions occur in toric colliders) and computational biology (cf. González-Sanz and Hundrieser (2023), where protein structures are recorded with pairs of dihedral angles). More generally, toric data arises in a variety of applications in directional statistics (e.g. Klein et al. (2020), Wiechers et al. (2023), etc.), and our results are naturally applicable to such settings.

We also note that periodicity constraints are commonly imposed in nonparametric estimation problems to mitigate boundary issues (Efromovich, 1999; Krishnamurthy et al., 2014; Han et al., 2020). In many such cases, an alternative is to assume that the underlying probability measures place sufficiently small mass near the boundary. Such an assumption cannot be used in our context since, as before, we shall require all densities to be bounded away from zero throughout their support. Optimal estimation rates under Wasserstein distances differ dramatically in the absence of a density lower bound condition (Bobkov and Ledoux, 2019; Niles-Weed and Berthet, 2022), and we do not address this setting here.

We now turn to our main results. Recall the background on the quadratic optimal transport problem over $\mathbb{T}^d$ in Section 1.3.3. Let $P, Q \in \mathcal{P}_{\mathrm{ac}}(\mathbb{T}^d)$ be absolutely continuous measures admitting respective $\mathbb{Z}^d$-periodic densities p and q. We now denote by T_0 the optimal transport map from P to Q, with respect to the cost $d_{\mathbb{T}^d}^2$. As outlined in Proposition 1, T_0 is the gradient of a convex potential $\varphi_0 : \mathbb{R}^d \to \mathbb{R}$, and is uniquely determined P-almost everywhere. We continue to denote by $\phi_0 = \|\cdot\|^2 - 2\varphi_0$ and $\psi_0 = \|\cdot\|^2 - 2\varphi_0^*$ a corresponding pair of Kantorovich potentials. Let $X_1, \dots, X_n \sim P$ and $Y_1, \dots, Y_m \sim Q$ denote i.i.d. samples, which are independent of each other, and let $\widehat{P}_n, \widehat{Q}_m$ respectively denote the distributions induced by density estimators $\widehat{p}_n, \widehat{q}_m$ of p, q over $\mathbb{T}^d$, to be defined below. Our aim is to bound the risk of the estimator

$$\widehat{T}_{nm} = \nabla\widehat{\varphi}_{nm} = \operatorname*{argmin}_{T\in\mathcal{T}(\widehat{P}_n,\widehat{Q}_m)} \int d^2_{\mathbb{T}^d}(T(x),x)d\widehat{P}_n(x). \tag{5.25}$$

Note that $\widehat{P}_n$ and $\widehat{Q}_m$ are absolutely continuous, thus there indeed exists a unique solution to the above minimization problem, by Proposition 1. We continue to quantify the risk of $\widehat{T}_{nm}$ in terms of the $L^2(P)$ loss

$$\big\|\widehat{T}_{nm} - T_0\big\|^2_{L^2(P)} = \int_{\mathbb{T}^d} \big\|\widehat{T}_{nm}(x) - T_0(x)\big\|^2 dP(x).$$

Notice that the integrand on the right-hand side of the above display is $\mathbb{Z}^d$-periodic by Proposition 3(ii) and by the optimality of $\widehat{T}_{nm}$ and T_0, thus it indeed defines a map $\mathbb{T}^d \to \mathbb{R}$. As before, we shall also obtain upper bounds on the bias and risk of $\mathcal{W}_2^2(\widehat{P}_n,\widehat{Q}_m)$ as a byproduct of our proofs. Indeed, our main results hinge upon the stability bounds derived in previous sections, which can easily be shown to hold in the present context.

Proposition 27. Assume φ_0 satisfies condition **A1(λ)**, in the sense that φ_0 is a twice differentiable convex function over $\mathbb{R}^d$ satisfying $\lambda^{-1}I_d \preceq \nabla^2\varphi_0(x) \preceq \lambda I_d$ for all $x \in \mathbb{R}^d$. Then, Theorem 18 and Proposition 23 hold with $\Omega = \mathbb{T}^d$.

We now turn to the choice of density estimators $(\widehat{p}_n, \widehat{q}_m)$. The absence of a boundary on the sampling domain $\mathbb{T}^d$ facilitates the analysis of kernel density estimation, which will be our main focus in this section. We also study periodic wavelet density estimators, similarly to the one-sample case, but we defer this analysis to Appendix 5.G in the interest of brevity.

Given a kernel $K \in \mathcal{C}_c^\infty(\mathbb{R}^d)$ and a bandwidth $h_n > 0$, write $K_{h_n} = h_n^{-d}K(\cdot/h_n)$, and define the kernel density estimators of p and q by

$$\widetilde{p}_n^{(\mathrm{ker})} = P_n \star K_{h_n} = \int_{\mathbb{R}^d} K_{h_n}(\cdot - z)dP_n(z), \quad \widetilde{q}_m^{(\mathrm{ker})} = Q_m \star K_{h_m} = \int_{\mathbb{R}^d} K_{h_m}(\cdot - z)dQ_m(z).$$

Recall that integration over $\mathbb{R}^d$ with respect to a measure in $\mathcal{P}(\mathbb{T}^d)$ is understood as integration with respect to this measure extended to $\mathbb{R}^d$ via translation by $\mathbb{Z}^d$-periodicity. The above estimators may take on negative values, thus we again define the final density estimators by

$$\widehat{p}_n^{(\mathrm{ker})} \propto \widetilde{p}_n^{(\mathrm{ker})} I(\widetilde{p}_n^{(\mathrm{ker})} \geq 0), \quad \widehat{q}_m^{(\mathrm{ker})} \propto \widetilde{q}_m^{(\mathrm{ker})} I(\widetilde{q}_m^{(\mathrm{ker})} \geq 0),$$

where the proportionality constants are to be chosen such that $\widehat{p}_n^{(\mathrm{ker})}$ and $\widehat{q}_m^{(\mathrm{ker})}$ are densities. We also denote their induced probability distributions by $\widehat{P}_n^{(\mathrm{ker})}$ and $\widehat{Q}_m^{(\mathrm{ker})}$. Furthermore, $\widehat{T}_{nm}^{(\mathrm{ker})}$ denotes the optimal transport map between these measures.

We shall require the following condition on the kernel K, for given real numbers $\zeta, \kappa > 0$.

K(α). $K \in \mathcal{C}_c^\infty((0,1)^d)$ is an even kernel which satisfies

$$\sup_{\xi\in\mathbb{R}^d\setminus\{0\}} \frac{|\mathcal{F}[K](\xi) - 1|}{\|\xi\|^\alpha} < \infty. \tag{5.26}$$

A sufficient condition for equation (5.26) to hold is for $K \in \mathcal{C}_c^\infty(\mathbb{R}^d)$ to be a kernel of order $\beta = \lceil \zeta - 1 \rceil$. Such a statement appears for instance in Tsybakov (2008) when $d = 1$, and can easily be generalized to $d > 1$. Multivariate kernels of order β which additionally lie in $\mathcal{C}_c^\infty(\mathbb{R}^d)$ can readily be defined; for example, one may start with a univariate even kernel $K_0 \in \mathcal{C}_c^\infty(\mathbb{R})$ of order β, constructed for instance using the procedure of Fan and Hu (1992), and then set $K(x) = \prod_{i=1}^d K_0(x_i)$ (Giné and Nickl, 2016).

Divol (2021) stated that their work may be used to show that $\widehat{P}_n^{(\mathrm{ker})}$ achieves a comparable rate of convergence as the boundary-corrected wavelet estimator $\widehat{P}_n^{(\mathrm{bc})}$, in Wasserstein distance. We provide a formal statement and proof of this fact in Lemma 46 of Appendix 5.H, and use it to derive the following result.

Theorem 20 (Kernel Estimators)**.** Let the distributions $P, Q \in \mathcal{P}_{\mathrm{ac}}(\mathbb{T}^d)$ admit densities $p, q \in \mathcal{C}^{\alpha-1}(\mathbb{T}^d; M, \gamma)$ for some $\alpha > 1$ and $M, \gamma > 0$. Assume further that K is a kernel satisfying condition **K(2$\boldsymbol{\alpha}$)**. Let $h_n \asymp n^{-1/(d+2(\alpha-1))}$. Then, there exists a constant $C > 0$ depending only on K, M, γ, α such that the following statements hold.

(i) (Optimal Transport Maps) We have,

$$\mathbb{E}\big\|\widehat{T}_{nm}^{(\mathrm{ker})} - T_0\big\|_{L^2(P)}^2 \leq C R_{K,n\wedge m}(\alpha), \quad \text{where } R_{K,n}(\alpha) := \begin{cases} n^{-\frac{2\alpha}{2(\alpha-1)+d}}, & d \geq 3 \\ \log n/n, & d = 2 \\ 1/n, & d = 1. \end{cases}$$

(ii) (Wasserstein Distances) Assume further that $\alpha \notin \mathbb{N}$. Then,

$$\begin{aligned} &\big|\mathbb{E}\mathcal{W}_2^2(\widehat{P}_n^{(\mathrm{ker})}, \widehat{Q}_m^{(\mathrm{ker})}) - \mathcal{W}_2^2(P, Q)\big| \leq C R_{K,n\wedge m}(\alpha), \\ &\mathbb{E}\big|\mathcal{W}_2^2(\widehat{P}_n^{(\mathrm{ker})}, \widehat{Q}_m^{(\mathrm{ker})}) - \mathcal{W}_2^2(P, Q)\big|^2 \leq \left[C R_{K,n\wedge m}(\alpha) + \sqrt{\frac{\mathrm{Var}_P[\phi_0(X)]}{n} + \frac{\mathrm{Var}_Q[\psi_0(Y)]}{m}}\right]^2. \end{aligned}$$

Theorem 20 shows that the plugin estimators $\widehat{T}_{nm}$ and $\mathcal{W}_2^2(\widehat{P}_n, \widehat{Q}_m)$ achieve similar convergence rates as in the one-sample setting of Theorem 19. Unlike the latter result, we also note that Theorem 20 places no conditions on the regularity of T_0 or φ_0. Indeed, over $\mathbb{T}^d$, these can be inferred from the assumption $p, q \in \mathcal{C}^{\alpha-1}(\mathbb{T}^d; M, \gamma)$, due to Theorem 4. We exclude the case $\alpha \in \mathbb{N}$ from Theorem 20(ii) due in part to our use of this result. Nevertheless, even when $\alpha \in \mathbb{N}$, Theorem 20(ii) implies that

$$\big|\mathbb{E}\mathcal{W}_2^2(\widehat{P}_n, \widehat{Q}_m) - \mathcal{W}_2^2(P, Q)\big| \lesssim_\epsilon R_{K,n\wedge m}^{1-\epsilon}(\alpha),$$

for any $\epsilon > 0$, and similarly for the risk of $\mathcal{W}_2^2(\widehat{P}_n, \widehat{Q}_m)$.

If one is willing to place assumptions on the regularity of the potentials φ_0 and φ_0^*, then an analogue of Theorem 20(ii) can be derived when the periodicity assumption is removed, and the sampling domain is simply the unit cube $[0, 1]^d$. Such a result is stated in Proposition 28 of Appendix 5.G, and is made possible by the fact that Proposition 23 does not require any regularity of the fitted potentials.

When $d = 2$, Theorem 20(i) exhibits an improved convergence rate relative to Theorem 19(i), scaling as $\log n/n$ instead of $(\log n)^2/n$, which we now briefly discuss. This rate arises from our upper bound on $\mathbb{E}\mathcal{W}_2^2(\widehat{P}_n^{(\mathrm{ker})}, P)$ in Lemma 46, which makes use of the inequality (5.9) comparing $\mathcal{W}_2$ to a negative-order homogeneous Sobolev norm (Peyre, 2018). This last implies

$$\mathcal{W}_2(\widehat{P}_n^{(\mathrm{ker})}, P) \lesssim \|\widehat{p}_n^{(\mathrm{ker})} - p\|_{\dot{H}^{-1}(\mathbb{T}^d)} \asymp \|\widehat{p}_n^{(\mathrm{ker})} - p\|_{\mathcal{B}^{-1}_{2,2}(\mathbb{T}^d)}. \tag{5.27}$$

In contrast, when $P \in \mathcal{P}_{\mathrm{ac}}([0,1]^d)$, our upper bounds for wavelet estimators (and implicitly for empirical estimators in Corollary 19) employed the following distinct relation, arising from the work of Niles-Weed and Berthet (2022),

$$W_2(\widehat{P}_n^{(\mathrm{bc})}, P) \lesssim \|\widehat{p}_n^{(\mathrm{bc})} - p\|_{\mathcal{B}^{-1}_{2,1}([0,1]^d)}, \tag{5.28}$$

and similarly for the estimator $\widehat{P}_n^{(\mathrm{per})}$ described in Appendix 5.G. It can be seen that the $\mathcal{B}^{-1}_{2,2}$ norm is weaker than the $\mathcal{B}^{-1}_{2,1}$ norm. While either of these norms provide sufficiently tight upper bounds in equations (5.27) and (5.28) to obtain the minimax rate for density estimation in Wasserstein distance when $d \neq 2$, the former allows for a tighter logarithmic factor to be derived when $d = 2$. Inspired by the celebrated Ajtai–Komlós–Tusnády matching theorem (Ajtai, Komlós, and Tusnády, 1984; Talagrand, 1992), it is natural to conjecture that the rate $\log n/n$ in the definition of $R_{K,n}(\alpha)$ cannot be further improved when $d = 2$, for any of the conclusions of Theorem 20.

Theorem 20 is proved in Appendix 5.H, where the main difficulty is to show that the evaluation $L(\widehat{Q}_m^{(\mathrm{ker})})$, of the linear functional L from equation (5.7), has bias decaying at the quadratic rate $h_m^{2\alpha}$. As for our analysis of wavelet estimators, this rate improves upon the naive upper bound $|\mathbb{E}L(\widehat{Q}_m^{(\mathrm{ker})})| \lesssim h_m^{\alpha-1}$, which could have been deduced from the traditional bias of kernel density estimators in estimating an $(\alpha-1)$-Hölder continuous density (Tsybakov, 2008). Similar considerations arise in the analysis of kernel-based estimators for other important functionals, such as the integral of a squared density (Giné and Nickl, 2008).

Remark 3 (Dependence Between Samples). Our assumption of independence between the sample points X_i and Y_j is only used to derive the sharp constant in the final term of Theorem 20(ii), which implies that, when $2(\alpha+1) > d$,

$$\mathbb{E}\big|W_2^2(\widehat{P}_n, \widehat{Q}_m) - W_2^2(P, Q)\big|^2 \leq (1 + o(1)) \left(\frac{\mathrm{Var}_P[\phi_0(X)]}{n} + \frac{\mathrm{Var}_Q[\psi_0(Y)]}{m}\right).$$

If one is willing to inflate these leading constants, then all assertions of Theorem 20 continue to hold under arbitrary dependence structures between the two i.i.d. samples.

5.3.4 Toward Two-Sample Estimation over Smooth Domains

Our aim is now to show that an analogue of Theorem 20 holds over generic domains $\Omega \subseteq \mathbb{R}^d$, provided that one is willing to assume that Caffarelli's global regularity theorem (Theorem 3(ii)) holds uniformly over Ω. Specifically, we will use the following conditions throughout this section.

- **(C1)** Ω is a known compact, convex subset of $\mathbb{R}^d$, such that $\partial\Omega$ is $\mathcal{C}^\infty$and $\mathcal{L}(\Omega) = 1$.
- **(C2)** There exists $\epsilon_0 > 0$ such that for any $M, \gamma > 0$ and $\epsilon \in (0, \epsilon_0)$, there exists a constant $C > 0$ depending only on $\Omega, M, \gamma, \epsilon$ such that for any densities $\widehat{p}, \widehat{q} \in \mathcal{C}^\epsilon(\Omega; M, \gamma)$, the unique mean-zero Brenier potential $\widehat{\varphi}$ whose gradient pushes forward $\widehat{p}$ onto $\widehat{q}$ satisfies
$$\|\widehat{\varphi}\|_{\mathcal{C}^2(\Omega)} \leq C.$$

As discussed previously, we are only able to verify condition **(C2)** when Ω is replaced by the torus $\mathbb{T}^d$. However, in view of Theorem 3, it is natural to conjecture that condition **(C2)** is satisfied for other domains of the type **(C1)**, and if such a result is proven in future work, then the bounds appearing in this section can be applied. For completeness, we will also state a one-sample result over Ω, for which condition **(C2)** is not needed.

It is well-known that kernel density estimators suffer from leading-order boundary bias, and are thus not minimax optimal for estimating strictly positive densities on compact subsets of $\mathbb{R}^d$. In order to develop a minimax optimal estimator for densities supported on Ω, we will impose Neumann boundary conditions on the densities, and we will introduce an orthonormal basis of $L^2(\Omega)$ generated by the eigenfunctions of the Neumann Laplacian. To elaborate, define $H^2_N(\Omega)$ to be the set of functions $u \in H^2(\Omega) \cap L^2_0(\Omega)$ satisfying
$$\frac{\partial u}{\partial \nu} = 0, \quad \text{over } \partial\Omega,$$
where ν is an outward-pointing normal vector to $\partial\Omega$, and the normal derivative is to be understood in the weak sense. Under condition **(C1)**, it is a standard fact that the negative Laplace operator $-\Delta$ is a self-adjoint bijection of $H^2_N(\Omega)$ onto $L^2_0(\Omega)$, which admits a real and discrete spectrum $0 < \lambda_1 \leq \lambda_2 \leq \dots$, with corresponding eigenfunctions $\{\eta_\ell\}_{\ell=1}^\infty \subseteq H^2_N(\Omega)$ (Dunlop et al., 2020; Evans, 1998). The latter form an orthonormal basis of $L^2_0(\Omega)$.

Let the distributions $P, Q \in \mathcal{P}_{\mathrm{ac}}(\Omega)$ admit densities $p, q \in L^2(\Omega)$, and let $X_1, \dots, X_n \sim P$ and $Y_1, \dots, Y_m \sim Q$ be i.i.d. observations. Under condition **(C1)**, the densities may be expanded as
$$p = 1 + \sum_{\ell=1}^\infty \alpha_\ell \eta_\ell, \quad q = 1 + \sum_{\ell=1}^\infty \beta_\ell \eta_\ell,$$
where $\alpha_\ell = \int \eta_\ell dP$ and $\beta_\ell = \int \eta_\ell dQ$. Let $\tau \in \mathcal{C}^\infty(\mathbb{R}_+)$ be a smooth approximation to the indicator function $I(\cdot < 1)$. Specifically, assume that τ is a nonincreasing and smooth function such that $\tau(x) = 1$ for all $x < 1/2$, and $\tau(x) = 0$ for all $x \geq 1$. Given an integer $L_n \geq 1$, set
$$\omega_\ell = \tau(\lambda_\ell / \lambda_{L_n}), \quad \ell = 1, 2, \dots,$$
and define the density estimators
$$\widetilde{p}_n^{(\mathrm{lap})} = 1 + \sum_{\ell=1}^{L_n} \omega_\ell \widehat{\alpha}_\ell \eta_\ell, \quad \widetilde{q}_m^{(\mathrm{lap})} = 1 + \sum_{\ell=1}^{L_m} \omega_\ell \widehat{\beta}_\ell \eta_\ell,$$

where $\widehat{\alpha}_\ell = \int \eta_\ell dP_n$, $\widehat{\beta}_\ell = \int \eta_\ell dQ_m$ for $\ell = 1, 2, \dots$. Density estimators of this type have also appeared in the works of Hendriks (1990) and Cleanthous et al. (2020). They may be thought of as truncated series estimators for which the truncation is smoothed by the weight function τ. As such, they are closely related to kernel density estimators. In fact, if one were to replace Ω by the torus $\mathbb{T}^d$ (in which case the Neumann boundary condition is replaced by the periodic boundary condition), then $\widetilde{p}_n^{(\mathrm{lap})}$ would precisely be the kernel density estimator whose kernel is the inverse Fourier transform of the map $\xi \in \mathbb{R}^d \mapsto \tau(\|2\pi\xi\|^2)$, and whose bandwidth is $\lambda_{L_n}^{-1/2}$. We also note that if one were to choose the nonsmooth function $\tau = I(\cdot < 1)$, then $\widetilde{p}_n^{(\mathrm{lap})}$ would reduce to a traditional series estimator, but our analysis does not extend to this case: the smoothness of τ is crucial for our use of $L^r(\Omega)$ multiplier arguments, as we discuss further in Remark 4 of Appendix 5.I.

As in previous sections, we define the final estimators to be the densities given by

$$\widehat{p}_n^{(\mathrm{lap})} \propto \widetilde{p}_n^{(\mathrm{lap})} I(\widetilde{p}_n^{(\mathrm{lap})} \geq 0), \quad \widehat{q}_m^{(\mathrm{lap})} \propto \widetilde{q}_m^{(\mathrm{lap})} I(\widetilde{q}_m^{(\mathrm{lap})} \geq 0),$$

and we let $\widehat{P}_n^{(\mathrm{lap})}$ and $\widehat{Q}_m^{(\mathrm{lap})}$ be the induced distributions. Furthermore, let $\widehat{T}_{nm}^{(\mathrm{lap})}$ be the optimal transport map from $\widehat{P}_n^{(\mathrm{lap})}$ to $\widehat{Q}_m^{(\mathrm{lap})}$, and $\overline{T}_m^{(\mathrm{lap})}$ the optimal transport map from P to $\widehat{Q}_m^{(\mathrm{lap})}$. We omit the superscripts “lap” for the remainder of this section. In order to state convergence rates for these estimators, we will work over the constrained Hölder spaces

$$\mathcal{C}_N^s(\Omega) = \left\{ u \in \mathcal{C}^s(\Omega) : \frac{\partial \Delta^j u}{\partial \nu} = 0 \text{ on } \partial\Omega,\ 0 \leq j \leq \left\lfloor \frac{s-1}{2} \right\rfloor \right\},$$

and the associated balls $\mathcal{C}_N^s(\Omega; M) = \mathcal{C}^s(\Omega; M) \cap \mathcal{C}_N^s(\Omega)$ and $\mathcal{C}_N^s(\Omega; M, \gamma) = \mathcal{C}^s(\Omega; M, \gamma) \cap \mathcal{C}_N^s(\Omega)$, for $M, \gamma, s > 0$. Note that $\mathcal{C}^s(\Omega) = \mathcal{C}_N^s(\Omega)$ for $s < 1$. Our main result is the following.

Theorem 21. Let Ω be a domain satisfying condition **(C1)**. Assume the distributions $P, Q \in \mathcal{P}_{\mathrm{ac}}(\Omega)$ admit densities $p, q \in \mathcal{C}_N^{\alpha-1}(\Omega; M, \gamma)$ for some $\alpha > 1$, $\alpha \notin \mathbb{N}$, and $M, \gamma > 0$. Let $L_n^{1/d} \asymp n^{1/(d+2(\alpha-1))}$. Then, there exists a constant $C > 0$ depending only on $\Omega, M, \gamma, \alpha, \tau$ such that the following statements hold.

(i) (One-Sample) Assume $\varphi_0 \in \mathcal{C}^2(\Omega; M)$. Then,

$$\mathbb{E}\big\|\overline{T}_m - T_0\big\|_{L^2(P)}^2 \leq C R_{K,m}(\alpha).$$

(ii) (Two-Sample) Assume that condition **(C2)** holds. Then,

$$\mathbb{E}\big\|\widehat{T}_{nm} - T_0\big\|_{L^2(P)}^2 \leq C R_{K,n\wedge m}(\alpha),$$

Under the strong condition **(C2)**, Theorem 21(ii) shows that our two-sample results on transport map estimation over the torus can be extended to generic domains[3] Ω, provided

[3] While we assume for simplicity that P and Q share the same support Ω, Theorem 21 can readily be extended to the case where P and Q are supported on distinct domains which both satisfy condition **(C1)**, with natural modifications to the statement of condition **(C2)** and to the definitions of $\widehat{p}_n$, $\widehat{q}_m$.

that one is willing to place Neumann boundary conditions on the true densities. While other boundary conditions may have been used in our analysis, it is important that they be chosen such that p, q are permitted to be smooth and strictly positive over Ω; in particular, one cannot impose the Dirichlet condition $p = q = 0$ over $\partial\Omega$. Our current boundary conditions are satisfied by a wide range of densities, such as those whose gradient vanishes at the boundary, or those which are equal to finite linear combinations of the eigenfunctions $\{\eta_\ell\}_{\ell=1}^\infty$. We also emphasize that Theorem 21 imposes no boundary conditions when $\alpha < 2$.

To prove Theorem 21, our primary contribution is to derive Propositions 33 and 34 of Appendix 5.I, which state convergence rates for $\widehat{p}_n$ under the spectral Sobolev norms $\mathcal{H}^{t,r}(\Omega)$ which we define therein. Here, $t \in \mathbb{R}$ and $r > 1$ are smoothness and integrability indices, respectively. Under some conditions on r, our results imply that for large enough d, and $-\infty < t < \alpha - 1$, it holds that

$$\mathbb{E}\|\widehat{p}_n - p\|_{\mathcal{H}^{t,r}(\Omega)} \lesssim n^{-\frac{\alpha-1-t}{2(\alpha-1)+d}}, \tag{5.29}$$

assuming the same conditions as Theorem 21. This bound has two implications:

- On the one hand, taking $t = -1$, $r = 2$, and applying equation (5.9), we deduce that $\widehat{p}_n$ is a minimax optimal density estimator under the 2-Wasserstein distance. To the best of our knowledge, this is the only known convergence rate for density estimation under the Wasserstein distance over Euclidean domains with non-rectangular boundary (apart from the special case $\alpha \in (1, 2]$, for which density estimation has been studied without support assumptions by Niles-Weed and Berthet (2022)). We also highlight that Divol, Niles-Weed, and Pooladian (2022) has studied the case of boundary-free manifolds.
- On the other hand, take $s = \epsilon/2$ for some $\epsilon > 0$, and let $r > 2d/\epsilon$. Then, using a Sobolev embedding argument, equation (5.29) leads to a convergence rate for $\widehat{p}_n$ under the $\mathcal{C}^\epsilon(\Omega)$ norm. In particular, this allows us to infer that $\widehat{p}_n$ and $\widehat{q}_m$ satisfy the conditions on the densities in assumption **(C2)**, with high probability, thus allowing us to infer that $\widehat{\varphi}_{nm}$ is of class $\mathcal{C}^{2+\epsilon}(\Omega)$ with uniformly bounded Hölder norm.

We close this section by noting that, if one is willing to settle for pointwise asympotics, then Theorem 21(i) can be stated without any smoothness assumptions on the potential φ_0. Indeed, the following is a consequence of Caffarelli's regularity theory (Theorem 3).

Corollary 21. *Let Ω be a domain satisfying condition **(C1)**. Assume the distributions $P, Q \in \mathcal{P}_{\mathrm{ac}}(\Omega)$ admit densities $p, q \in \mathcal{C}_N^{\alpha-1}(\Omega)$ for some $\alpha > 1$. Let $L_m^{1/d} \asymp m^{1/(d+2(\alpha-1))}$. Then, there exists a constant $C > 0$ depending on $\Omega, \alpha, \tau, p, q$ such that*

$$\mathbb{E}\big\|\overline{T}_m - T_0\big\|_{L^2(P)}^2 \leq C R_{K,m}(\alpha).$$

5.4 Discussion

We have shown that several families of plugin estimators for smooth optimal transport maps are minimax optimal. Our analysis hinged upon stability arguments which relate this problem to that of estimating the Wasserstein distance between two distributions, and, in turn, to that of estimating a distribution under the Wasserstein distance. The latter question is well-studied in the literature, and formed a key component in deriving convergence rates for the former two problems. As a byproduct of our stability results, we derived new upper bounds on the convergence rate of estimators of the Wasserstein distance between sufficiently smooth distributions, which could not have been deduced from our developments in Chapter 3.

The estimators in this work are simple to compute and minimax optimal, but we make no claim that their computational efficiency is optimal. For example, our plugin estimators of the Wasserstein distance between $(\alpha - 1)$-smooth densities can be approximated by sampling N observations from our density estimators, and computing the Wasserstein distance between the empirical measures formed by these observations, which can be done in polynomial time with respect to N (Peyré and Cuturi, 2019). In order for this approximation to achieve comparable risk to our theoretical estimators in the high-smoothness regime $\alpha \gtrsim d$, one must take $N \asymp n^{cd}$ for some $c \geq 1$. Our estimator thus requires computation time depending exponentially on d. Vacher et al. (2021, 2024); Lin, Cuturi, and Jordan (2024) analyzed alternative estimators based on kernel sum-of-squares, which have more favorable computational properties; though their estimators are not minimax optimal, they can be computed in polynomial time if $\alpha \gtrsim d$. It is an interesting open question to derive polynomial-time estimators in d which are also minimax optimal. More broadly, there are other computationally efficient estimators for optimal transport maps based on entropic regularization (Cuturi, 2013) and input convex neural networks (Makkuva et al., 2020) whose L^2 risks have very recently been studied (Pooladian and Niles-Weed, 2021; Divol, Niles-Weed, and Pooladian, 2022), but are not yet known to achieve the minimax rate.

In our analysis of smooth two-sample optimal transport map estimators, we required the fitted Brenier potential to be twice Hölder-smooth, for which we appealed to Caffarelli's regularity theory. Since we do not know whether Caffarelli's boundary regularity estimates hold uniformly in the various problem parameters, we resorted to working over $\mathbb{T}^d$, where a uniform analogue of Caffarelli's theory is available (cf. Theorem 4). We showed in Section 5.3.4 that this analysis can be extended to generic domains Ω of $\mathbb{R}^d$, conditionally on a uniform version of Caffarelli's boundary regularity theory. To the best of our knowledge, it remains an interesting open to question to verify whether this condition indeed holds.

Finally, our work leaves open the question of estimating optimal transport maps when the ground cost function is not the squared Euclidean norm. While each of the plugin estimators in this chapter can be naturally defined for generic cost functions, their theoretical analysis presents a breadth of challenges. For example, although the regularity theory of Caffarelli has been generalized to cover a large collection of cost functions (Ma, Trudinger, and Wang, 2005), this collection does not include the costs $\|\cdot\|^p$ for $p \neq 2$ and $p > 1$, which are arguably most widely-used in statistical applications. For such costs, it remains unclear what regularity conditions are sensible to place on the population optimal transport map in order to obtain analogues of our risk bounds, and we hope to explore such questions in future work.

5.A On the Variance of Kantorovich Potentials

We state a straightforward technical result which will be used throughout our proofs.

Lemma 37. *Let Ω be equal to $[0,1]^d$ or $\mathbb{T}^d$. Given $P, Q \in \mathcal{P}_{\mathrm{ac}}(\Omega)$, let (ϕ_0, ψ_0) be a pair of Kantorovich potentials in the optimal transport transport problem from P to Q. Assume further that the density q of Q satisfies $\gamma^{-1} \leq q \leq \gamma$ over Ω, for some $\gamma > 0$. Define $\bar{\psi}_0 = \psi_0 - \int_\Omega \psi_0$. Then, there exists a constant $C > 0$ depending only on d such that*

$$\|\bar{\psi}_0\|_{L^2(Q)} \leq C\gamma W_2(P,Q).$$

In particular,

$$\mathrm{Var}_Q[\psi_0(Y)] \leq (C\gamma)^2 W_2^2(P,Q).$$

The proof will follow from Poincaré inequalities over $[0,1]^d$ and $\mathbb{T}^d$, which we recall here as they will be needed again in the sequel. The following is a special case of the Poincaré inequality for convex domains (see for instance Leoni (2017), Theorem 12.30).

Lemma 38. *Let $0 < a < b < \infty$ and $\Omega = [a,b]^d$. Then, there exists a constant $C > 0$ depending only on d such that for all $f \in H^1(\Omega)$ satisfying $\int_\Omega f = 0$,*

$$\|f\|_{L^2(\Omega)} \leq C(b-a) \|\nabla f\|_{L^2(\Omega)}.$$

We also state the following classical periodic Poincaré inequality (see for instance Steinerberger (2016) for a simple proof).

Lemma 39. *Let $f \in H^1(\mathbb{T}^d)$ satisfy $\int_{\mathbb{T}^d} f = 0$. Then, $\|f\|_{L^2(\mathbb{T}^d)} \leq \|\nabla f\|_{L^2(\mathbb{T}^d)}$.*

Proof of Lemma 37. Since $\psi_0 \in H^1(\Omega)$ by definition, we may apply the Poincaré inequality over Ω (namely, Lemma 38 when $\Omega = [0,1]^d$, or Lemma 39 when $\Omega = \mathbb{T}^d$). This fact, together with the assumption $\gamma^{-1} \leq q \leq \gamma$, implies

$$\|\bar{\psi}_0\|_{L^2(Q)}^2 \leq \gamma \|\bar{\psi}_0\|_{L^2([0,1]^d)}^2 \leq C^2\gamma \|\nabla\psi_0\|_{L^2([0,1]^d)}^2 \leq (C\gamma)^2 \|\nabla\psi_0\|_{L^2(Q)}^2 = (C\gamma)^2 W_2^2(P,Q),$$

which then also implies

$$\mathrm{Var}_Q[\psi_0(Y)] = \mathrm{Var}_Q[\bar{\psi}_0(Y)] \leq \|\bar{\psi}_0\|_{L^2(Q)}^2 \leq (C\gamma)^2 W_2^2(P,Q),$$

as claimed. □

5.B Proofs of One-Sample Stability Bounds

5.B.1 Proof of Theorem 18

Recall that φ_0 denotes a Brenier potential from P to Q, while $\phi_0 = \|\cdot\|^2 - 2\varphi_0$ and $\psi_0 = \|\cdot\|^2 - 2\varphi_0^*$ denote the corresponding Kantorovich potentials. Since we have assumed that both

P and Q are absolutely continuous distributions, Brenier's Theorem implies that $S_0 = \nabla\varphi_0^*$ is the optimal transport map from Q to P. Since φ_0 is closed, the assumption

$$\frac{1}{\lambda}I_d \preceq \nabla^2\varphi_0 \preceq \lambda I_d,$$

from condition **A1(λ)** also implies (Hiriart-Urruty and Lemaréchal (2004), Theorem 4.2.2),

$$\frac{1}{\lambda}I_d \preceq \nabla^2\varphi_0^* \preceq \lambda I_d.$$

Combining this bound with a second-order Taylor expansion of φ_0^* leads to the following inequalities

$$\frac{1}{2\lambda}\left\|x-y\right\|^2 \leq \varphi_0^*(y) - \varphi_0^*(x) - \big\langle S_0(x), y-x\big\rangle \leq \frac{\lambda}{2}\left\|x-y\right\|^2, \quad x,y\in\Omega. \tag{5.30}$$

With these facts in place, we turn to proving the theorem, namely that

$$\frac{1}{\lambda}\|\widehat{T}-T_0\|_{L^2(P)}^2 \leq W_2^2(P,\widehat{Q}) - W_2^2(P,Q) - \int \psi_0 d(\widehat{Q}-Q) \leq \lambda W_2^2(\widehat{Q},Q). \tag{5.31}$$

We begin with the first inequality. Since $\widehat{T}$ is the optimal transport map from P to $\widehat{Q}$, we have,

$$\begin{aligned}
W_2^2(P,\widehat{Q}) &= \int \|\widehat{T}(x)-x\|^2 dP(x)\\
&= \int \|T_0(x)-x\|^2 dP(x)\\
&+ \int 2\big\langle T_0(x)-x, \widehat{T}(x)-T_0(x)\big\rangle dP(x) + \int \|\widehat{T}(x)-T_0(x)\|^2 dP(x)\\
&= W_2^2(P,Q) + \int 2\big\langle T_0(x)-x, \widehat{T}(x)-T_0(x)\big\rangle dP(x) + \|\widehat{T}-T_0\|_{L^2(P)}^2.
\end{aligned}$$

To bound the cross term, notice that equation (5.30) implies

$$\begin{aligned}
&2\int \big\langle T_0(x)-x, \widehat{T}(x)-T_0(x)\big\rangle dP(x)\\
&\quad= 2\int \big\langle T_0(x)-S_0(T_0(x)), \widehat{T}(x)-T_0(x)\big\rangle dP(x)\\
&\quad\geq 2\int \Big[\big\langle T_0(x), \widehat{T}(x)-T_0(x)\big\rangle\\
&\qquad\qquad + \varphi_0^*(T_0(x)) - \varphi_0^*(\widehat{T}(x)) + \frac{1}{2\lambda}\|\widehat{T}(x)-T_0(x)\|^2\Big] dP(x)\\
&\quad= \int \Big[\|\widehat{T}(x)\|^2 - \|T_0(x)\|^2 - \|\widehat{T}(x)-T_0(x)\|^2\\
&\qquad\qquad + 2\varphi_0^*(T_0(x)) - 2\varphi_0^*(\widehat{T}(x)) + \frac{1}{\lambda}\|\widehat{T}(x)-T_0(x)\|^2\Big] dP(x)
\end{aligned}$$

$$= \left(\frac{1}{\lambda} - 1\right) \|\widehat{T} - T_0\|_{L^2(P)}^2 + \int \psi_0 d(\widehat{Q} - Q).$$

We deduce

$$W_2^2(P, \widehat{Q}) \geq W_2^2(P, Q) + \frac{1}{\lambda}\|\widehat{T} - T_0\|_{L^2(P)}^2 + \int \psi_0 d(\widehat{Q} - Q),$$

To prove the second inequality in equation (5.31), let $\widehat{\pi}$ denote an optimal coupling between Q and $\widehat{Q}$. Then, the measure $\widehat{\pi}_{S_0} = (S_0, Id)_{\#}\widehat{\pi}$ is a (possibly suboptimal) coupling between P and $\widehat{Q}$, thus

$$W_2^2(P, \widehat{Q}) \leq \int \|x - z\|^2 \, d\widehat{\pi}_{S_0}(x, z) = \int \|S_0(y) - z\|^2 \, d\widehat{\pi}(y, z). \tag{5.32}$$

The claim is now a consequence of the following technical Lemma, which will be used again in the sequel.

Lemma 40. *We have,*

$$W_2^2(P, \widehat{Q}) \leq \int \|S_0(y) - z\|^2 \, d\widehat{\pi}(y, z) \leq W_2^2(P, Q) + \int \psi_0 d(\widehat{Q} - Q) + \lambda W_2^2(\widehat{Q}, Q).$$

5.B.2 Proof of Lemma 40

We have,

$$\begin{aligned}
&\int \|S_0(y) - z\|^2 \, d\widehat{\pi}(y, z) \\
&\quad = \int \|S_0(y) - y\|^2 \, dQ(y) + \int \|y - z\|^2 \, d\widehat{\pi}(y, z) + 2\int \big\langle S_0(y) - y, y - z\big\rangle d\widehat{\pi}(y, z) \\
&\quad = W_2^2(P, Q) + W_2^2(\widehat{Q}, Q) + 2\int \big\langle S_0(y) - y, y - z\big\rangle d\widehat{\pi}(y, z).
\end{aligned}$$

Now, notice that by (5.30),

$$\begin{aligned}
2\int \big\langle S_0(y), y - z\big\rangle d\widehat{\pi}(y, z) &\leq 2\int \left[\varphi_0^*(y) - \varphi_0^*(z) + \frac{\lambda}{2}\|y - z\|^2\right] d\widehat{\pi}(y, z) \\
&= 2\int \varphi_0^* d(Q - \widehat{Q}) + \lambda W_2^2(\widehat{Q}, Q),
\end{aligned}$$

and,

$$\begin{aligned}
&2\int \big\langle -y, y - z\big\rangle d\widehat{\pi}(y, z) \\
&\qquad = \int \left[\|z\|^2 - \|z - y\|^2 - \|y\|^2\right] d\widehat{\pi}(y, z) = \int \|\cdot\|^2 \, d(\widehat{Q} - Q) - W_2^2(\widehat{Q}, Q).
\end{aligned}$$

Therefore,

$$\begin{aligned}
&W_2^2(P, \widehat{Q}) - W_2^2(P, Q) \\
&\qquad \leq \int \left(\|\cdot\|^2 - 2\varphi_0^*\right) d(\widehat{Q} - Q) + \lambda W_2^2(\widehat{Q}, Q) = \int \psi_0 d(\widehat{Q} - Q) + \lambda W_2^2(\widehat{Q}, Q),
\end{aligned}$$

and the claim follows. □

5.C Proofs of Upper Bounds for One-Sample Empirical Estimators

In this Appendix, we prove Corollaries 19 and 20.

5.C.1 Proof of Corollary 19

We shall make use of the notation introduced in Section 5.2.3 and Appendix A.3.4, regarding wavelet density estimation over $[0,1]^d$. In particular, let $\Psi = \Psi^{\mathrm{bc}}$ with $N = 1$, so that Ψ is the Haar wavelet basis on $[0,1]^d$.

Lemma 41. *Let $J \geq 1$ be an integer. For any $\mu \in \mathcal{P}([0,1]^d)$, let $\mu_J \in \mathcal{P}_{\mathrm{ac}}([0,1]^d)$ denote the measure admitting density*

$$q_J = 1 + \sum_{j=0}^{J} \sum_{\xi \in \Psi_j} \xi \int \xi d\mu,$$

with respect to the Lebesgue measure on $[0,1]^d$. Then, $W_2(\mu, \mu_J) \leq \sqrt{d} 2^{-J}$.

Proof. The Lemma is a consequence of dyadic partitioning arguments which have previously been used by Boissard and Le Gouic (2014); Fournier and Guillin (2015); Weed and Bach (2019); Lei (2020). In particular, for all $j \geq 0$, let $\mathcal{Q}_j$ denote the natural partition (up to intersections on Lebesgue null sets) of $[0,1]^d$ into 2^{dj} cubes of length 2^{-j}. Then, Proposition 1 of Weed and Bach (2019) implies

$$W_2^2(\mu, \mu_J) \leq d \left[2^{-2J} + \sum_{j=1}^{J} 2^{-2(j-1)} \sum_{S \in \mathcal{Q}_j} |\mu(S) - \mu_J(S)| \right].$$

To prove the claim, it thus suffices to show that $\mu(S) = \mu_J(S)$ for all $S \in \mathcal{Q}_j$ and $j = 1, \dots, J$.

Let $j \geq 0$, $S \in \mathcal{Q}_j$, and recall that I_S is the indicator function of S. Denote its expansion in the Haar basis by

$$I_S = \mathcal{L}(S) + \sum_{\ell=0}^{\infty} \sum_{\xi \in \Psi_\ell} \gamma_\xi \xi, \quad \text{where } \gamma_\psi = \int I_S \psi, \psi \in \Psi.$$

Notice that for any $\ell \geq j$ and $\xi \in \Psi_\ell$, we have

$$\mathrm{supp}(\xi) \subseteq I_S, \quad \text{or} \quad \mathrm{supp}(\xi) \cap I_S = \emptyset.$$

Furthermore, since $\zeta = I_{[0,1]^d}$, and the Haar basis is orthonormal, we must have $\int_{[0,1]^d} \xi = 0$ for any $\xi \in \Psi_j$, $j \geq 0$. It must follow that

$$\gamma_\xi = \int I_S \xi = 0, \quad \text{for all } \xi \in \Psi_\ell, \ell \geq j,$$

that is, $I_S \in \text{Span}\left(\Phi \cup \bigcup_{\ell=0}^{j-1} \Psi_\ell\right)$. We therefore have, for any $S \in \mathcal{Q}_j$ and $j \leq J$,

$$\begin{aligned}
\mu_J(S) &= \int I_S(y) q_J(y) dy \\
&= \mathcal{L}(S) + \sum_{j=0}^{J} \sum_{\xi \in \Psi_j} \left(\int \xi d\mu\right) \left(\int I_S(y)\xi(y)dy\right) \\
&= \mathcal{L}(S) + \sum_{j=0}^{J} \sum_{\xi \in \Psi_j} \left(\int \xi d\mu\right) \gamma_\xi \\
&= \int \left(\mathcal{L}(S) + \sum_{j=0}^{J} \sum_{\xi \in \Psi_j} \xi \gamma_\xi\right) d\mu = \int I_S d\mu = \mu(S).
\end{aligned}$$

The claim follows. □

To prove the Corollary from here, let $2^{J_n} \asymp n^{1/d}$, and let $\widehat{Q}_n$ be the distribution with density

$$\widehat{q}_n(y) = 1 + \sum_{j=0}^{J_n} \sum_{\xi \in \Psi_j} \left(\int \xi dQ_n\right) \xi(y), \quad y \in [0,1]^d.$$

Apply Lemma 41 to the measure $\mu = Q_n$ to obtain

$$W_2^2(Q_n, Q) \lesssim W_2^2(Q_n, \widehat{Q}_n) + W_2^2(\widehat{Q}_n, Q) \lesssim 2^{-2J_n} + W_2^2(\widehat{Q}_n, Q) \lesssim n^{-2/d} + W_2^2(\widehat{Q}_n, Q).$$

Furthermore, recall that $\gamma^{-1} \leq q \leq \gamma$, thus we may apply Lemma 106 to deduce

$$\mathbb{E}W_2^2(\widehat{Q}_n, Q) \lesssim \begin{cases} n^{-2/d}, & d \geq 3 \\ (\log n)^2/n, & d = 2 \\ 1/n, & d = 1. \end{cases}$$

The claim follows. □

5.C.2 Proof of Corollary 20

By Theorem 18,

$$\mathbb{E}\left|W_2^2(P, Q_n) - W_2^2(P, Q)\right| \leq \mathbb{E}W_2^2(Q_n, Q) + \mathbb{E}\left|\int \psi_0 d(Q_n - Q)\right|.$$

By Jensen's inequality, the final term satisfies

$$\mathbb{E}\left|\int \psi_0 d(Q_n - Q)\right| \leq n^{-\frac{1}{2}} \sqrt{\text{Var}_Q[\psi_0(Y)]}.$$

Since ψ_0 is uniformly bounded by a constant depending only on d, the right-hand side of the above display is of order $n^{-1/2}$. Furthermore, by equation (5.12), we have $\mathbb{E}W_2^2(Q_n, Q) \lesssim \kappa_n$, thus the first part of the claim follows.

Under the assumptions of the second part of the claim, we may instead use Corollary 19 to obtain the stronger bound $\mathbb{E}W_2^2(Q_n, Q) \lesssim \bar{\kappa}_n$, as well as Lemma 37 to derive $\mathrm{Var}_Q[\psi_0(Y)] \lesssim W_2^2(P, Q)$. The claim then follows. □

5.D Proofs of Upper Bounds for One-Sample Wavelet Estimators

5.D.1 Proof of Theorem 19

Under the assumptions of part (i), we may apply Theorem 18 and Lemma 106 to obtain,

$$\mathbb{E}\big\|\widehat{T}_n - T_0\big\|_{L^2(P)}^2 \lesssim_\lambda \mathbb{E}W_2^2(\widehat{Q}_n, Q) \lesssim_{M,\gamma,\alpha} R_{T,n}(\alpha),$$

which immediately leads to the first claim. To prove the second claim, recall that we have assumed $\alpha > 1$, whence the assumption on φ_0^* implies in particular that $\|\varphi_0^*\|_{\mathcal{C}^2(\Omega)} \leq \lambda$. Since the densities p, q are bounded from below by γ^{-1} over $[0,1]^d$, and also bounded from above due to their Hölder continuity and the compactness of $[0,1]^d$, it follows by Lemma 35 that φ_0 satisfies condition **A1(λ)**, after possibly modifying the value of λ in terms of γ. We may therefore invoke Theorem 18 to obtain,

$$L(\widehat{Q}_n) \leq W_2^2(P, \widehat{Q}_n) - W_2^2(P, Q) \leq \lambda W_2^2(\widehat{Q}_n, Q) + L(\widehat{Q}_n).$$

Let $C > 0$ be a constant depending only on $M, \lambda, \gamma, \alpha$, whose value may change from line to line. By Lemma 106, we have

$$\mathbb{E}W_2^2(\widehat{Q}_n, Q) \leq C R_{T,n}(\alpha), \quad \text{and} \quad \mathbb{E}W_2^4(\widehat{Q}_n, Q) \leq C R_{T,n}^2(\alpha).$$

Furthermore, by Lemma 36, we have

$$\begin{aligned} \big|\mathbb{E}L(\widehat{Q}_n)\big| &\leq C R_{T,n}(\alpha) \\ \mathrm{Var}\left[L(\widehat{Q}_n)\right] &\leq \frac{1}{n}\left(\mathrm{Var}_Q[\psi_0(Y)] + 2^{-2J_n\alpha}\right) \leq \frac{\mathrm{Var}_Q[\psi_0(Y)]}{n} + C R_{T,n}^2(\alpha) \\ \mathbb{E}\big|L(\widehat{Q}_n)\big|^2 = \big|\mathbb{E}L(\widehat{Q}_n)\big|^2 + \mathrm{Var}\left[L(\widehat{Q}_n)\right] &\leq \frac{\mathrm{Var}_Q[\psi_0(Y)]}{n} + C R_{T,n}^2(\alpha). \end{aligned}$$

Combining the preceding three displays, we deduce that

$$\big|\mathbb{E}W_2^2(P, \widehat{Q}_n) - W_2^2(P, Q)\big| \leq \lambda \mathbb{E}W_2^2(\widehat{Q}_n, Q) + \big|\mathbb{E}L(\widehat{Q}_n)\big| \leq C R_{T,n}(\alpha), \tag{5.33}$$

and,

$$\mathbb{E}\big|W_2^2(P, \widehat{Q}_n) - W_2^2(P, Q)\big|^2$$

$$
\begin{aligned}
&\leq \mathbb{E}\left[\left(\lambda W_2^2(\widehat{Q}_n, Q) + \big|L(\widehat{Q}_n)\big|\right)^2\right] \\
&\leq \lambda^2 \mathbb{E} W_2^4(\widehat{Q}_n, Q) + 2\lambda \mathbb{E}\left[W_2^2(\widehat{Q}_n, Q)\big|L(\widehat{Q}_n)\big|\right] + \mathbb{E}\big|L(\widehat{Q}_n)\big|^2 \\
&\leq \lambda^2 \mathbb{E} W_2^4(\widehat{Q}_n, Q) + 2\lambda\sqrt{\left(\mathbb{E} W_2^4(\widehat{Q}_n, Q)\right)\mathbb{E}\big|L(\widehat{Q}_n)\big|^2} + \mathbb{E}\big|L(\widehat{Q}_n)\big|^2 \\
&\leq C\lambda^2 R_{T,n}^2(\alpha) + 2\lambda\sqrt{CR_{T,n}^2(\alpha)\left(CR_{T,n}^2(\alpha) + \frac{\mathrm{Var}_Q[\psi_0(Y)]}{n}\right)} + \frac{\mathrm{Var}_Q[\psi_0(Y)]}{n} \\
&\leq C^2 R_{T,n}^2(\alpha) + 2CR_{T,n}(\alpha)\sqrt{\frac{\mathrm{Var}_Q[\psi_0(Y)]}{n}} + \frac{\mathrm{Var}_Q[\psi_0(Y)]}{n} \\
&\leq \left(CR_{T,n}(\alpha) + \sqrt{\frac{\mathrm{Var}_Q[\psi_0(Y)]}{n}}\right)^2.
\end{aligned}
$$

The claim follows □

It thus remains to prove Lemma 36.

5.D.2 Proof of Lemma 36

In order to bound the bias of $\int \psi_0 \widehat{q}_n$, recall from Lemma 105 that the event $A_n = \{\widehat{q}_n = \widetilde{q}_n\}$ satisfies $\mathbb{P}(A_n^{\mathsf{c}}) \lesssim n^{-2}$. Since ψ_0 is bounded by a constant depending only on d, we have,

$$
\mathbb{E}\left|\int \psi_0(\widehat{q}_n - \widetilde{q}_n)\right| \leq \mathbb{E}\left(\left|\int \psi_0(\widehat{q}_n - \widetilde{q}_n)\right| I_{A_n^{\mathsf{c}}}\right) \lesssim \mathbb{P}(A_n^{\mathsf{c}}) \lesssim 1/n^2.
$$

We deduce that

$$
\big|\mathbb{E}\big[L(\widehat{Q}_n)\big]\big| \lesssim \left|\int \psi_0(\widetilde{q}_n - q)\right| + \frac{1}{n^2},
$$

thus we are left with bounding the bias of $\int \psi_0 \widetilde{q}_n$. Recall that $\widehat{\beta}_\xi$ is an unbiased estimator of β_ξ for all $\xi \in \Psi$, so that

$$
q_{J_n} := \mathbb{E}[\widetilde{q}_n] = \sum_{\zeta \in \Phi} \beta_\zeta \zeta + \sum_{j=j_0}^{J_n} \sum_{\xi \in \Psi_j} \beta_\xi \xi.
$$

Write the expansion of ψ_0 in the basis Ψ as

$$
\psi_0 = \sum_{\zeta \in \Phi} \gamma_\zeta \zeta + \sum_{j=j_0}^{\infty} \sum_{\xi \in \Psi_j} \gamma_\xi \xi, \quad \text{where } \gamma_\xi = \int \psi_0 \xi \text{ for all } \xi \in \Psi,
$$

where the series converges uniformly due to the Hölder regularity of ψ_0, so that,

$$
\int \psi_0(q - q_{J_n}) = \int \left(\sum_{\zeta \in \Phi} \gamma_\zeta \zeta + \sum_{j=j_0}^{\infty} \sum_{\xi \in \Psi_j} \gamma_\xi \xi\right)\left(\sum_{j=J_n+1}^{\infty} \sum_{\xi \in \Psi_j} \beta_\xi \xi\right) = \sum_{j=J_n+1}^{\infty} \sum_{\xi \in \Psi_j} \gamma_\xi \beta_\xi,
$$

by orthonormality of the basis Ψ. By Lemma 30(i) in Appendix A.3, we have $|\Psi_j| \lesssim 2^{dj}$, therefore

$$\left|\int \psi_0(q - q_{J_n})\right| \le \sum_{j=J_n+1}^{\infty} \sum_{\xi \in \Psi_j} |\gamma_\xi \beta_\xi| \lesssim \sum_{j=J_n+1}^{\infty} 2^{dj} \|(\gamma_\xi)_{\xi\in\Psi_j}\|_\infty \|(\beta_\xi)_{\xi\in\Psi_j}\|_\infty. \tag{5.34}$$

On the other hand, we have $\|\cdot\|_{\mathcal{B}^s_{\infty,\infty}(\Omega)} \lesssim \|\cdot\|_{\mathcal{C}^s(\Omega)}$ for all $s > 0$ by Lemma 102. Therefore, by assumption on q and φ_0^*, we obtain

$$\begin{aligned}
\|(\beta_\xi)_{\xi\in\Psi_j}\|_{\ell_\infty} &\le \|q\|_{\mathcal{B}^{\alpha-1}_{\infty,\infty}(\Omega)}\, 2^{-j[(\alpha-1)+\frac{d}{2}]} \lesssim 2^{-j[(\alpha-1)+\frac{d}{2}]}, \\
\|(\gamma_\xi)_{\xi\in\Psi_j}\|_{\ell_\infty} &\le \|\psi_0\|_{\mathcal{B}^{\alpha+1}_{\infty,\infty}(\Omega)}\, 2^{-j[(\alpha+1)+\frac{d}{2}]} \lesssim 2^{-j[(\alpha+1)+\frac{d}{2}]},
\end{aligned} \tag{5.35}$$

for all $j \ge j_0$. Combine equations (5.34)–(5.35) to deduce

$$\left|\mathbb{E}L(\widehat{Q}_n)\right| \lesssim \sum_{j=J_n+1}^{\infty} 2^{dj} 2^{-j[(\alpha+1)+\frac{d}{2}]} 2^{-j[(\alpha-1)+\frac{d}{2}]} \lesssim \sum_{j=J_n+1}^{\infty} 2^{-2j\alpha} \lesssim 2^{-2J_n\alpha} \asymp n^{-\frac{2\alpha}{2(\alpha-1)+d}}.$$

We next bound the variance $\mathrm{Var}_Q[L(\widehat{Q}_n)]$. Denote by

$$\psi_{J_n} = \sum_{\zeta\in\Phi} \gamma_\zeta \zeta + \sum_{j=j_0}^{J_n} \sum_{\xi\in\Psi_j} \xi \gamma_\xi$$

the projection of ψ_0 onto $\mathrm{Span}\left(\Phi \cup \bigcup_{j=j_0}^{J_n} \Psi_j\right)$. By again applying Lemma 105, it is a straightforward observation that

$$\left|\mathrm{Var}\left[\int \psi_0 \widehat{q}_n\right] - \mathrm{Var}\left[\int \psi_0 \widetilde{q}_n\right]\right| \lesssim n^{-2},$$

thus it suffices to show that $\mathrm{Var}\left[\int \psi_0 \widetilde{q}_n\right] = \mathrm{Var}_Q[\psi_0(Y)]/n + O(2^{-2J_n\alpha}/n)$. Notice that

$$\begin{aligned}
\int \psi_0 \widetilde{q}_n &= \sum_{\zeta\in\Phi} \widehat{\beta}_\zeta \int \psi_0 \zeta + \sum_{j=j_0}^{J_n} \sum_{\xi\in\Psi_j} \widehat{\beta}_\xi \int \psi_0 \xi \\
&= \sum_{\zeta\in\Phi} \widehat{\beta}_\zeta \gamma_\zeta + \sum_{j=j_0}^{J_n} \sum_{\xi\in\Psi_j} \widehat{\beta}_\xi \gamma_\xi \\
&= \frac{1}{n} \sum_{i=1}^{n} \left[\sum_{\zeta\in\Phi} \zeta(Y_i)\gamma_\zeta + \sum_{j=j_0}^{J_n} \sum_{\xi\in\Psi_j} \xi(Y_i)\gamma_\xi\right] = \int \psi_{J_n} dQ_n,
\end{aligned} \tag{5.36}$$

whence,

$$\mathrm{Var}\left[\int \psi_0 \widetilde{q}_n\right] = \frac{1}{n} \mathrm{Var}_Q[\psi_{J_n}(Y)]$$

$$= \frac{1}{n}\operatorname{Var}_Q[\psi_0(Y)] + \frac{1}{n}(\operatorname{Var}_Q[\psi_{J_n}(Y)] - \operatorname{Var}_Q[\psi_0(Y)]).$$

It thus remains to bound the final term. Notice that

$$\begin{aligned}&\Big|\operatorname{Var}_Q[\psi_{J_n}(Y)] - \operatorname{Var}_Q[\psi_0(Y)]\Big| \\ &\qquad\lesssim \Big|\mathbb{E}_Q[\psi_{J_n}^2(Y) - \psi_0^2(Y)]\Big| + \Big|\mathbb{E}_Q[\psi_{J_n}(Y) - \psi_0(Y)]\Big| = (I) + (II).\end{aligned}$$

We begin by bounding (I). Letting $g_n = (\psi_{J_n} + \psi_0)q$, we have,

$$(I) = \left|\int(\psi_{J_n} - \psi_0)(\psi_{J_n} + \psi_0)q\right| = \left|\int(\psi_{J_n} - \psi_0)g_n\right|.$$

It is clear that $\|\psi_{J_n}\|_{\mathcal{B}^{\alpha+1}_{\infty,\infty}([0,1]^d)} \leq \|\psi_0\|_{\mathcal{B}^{\alpha+1}_{\infty,\infty}([0,1]^d)} \lesssim \lambda$, thus for any fixed $\epsilon > 0$ sufficiently small, the map $\psi_{J_n} + \psi_0$ lies in $\mathcal{C}^{\alpha+1-\epsilon}([0,1]^d)$ with uniformly bounded norm, by Lemma 102. Note that one may take $\epsilon = 0$ if α is not an integer. On the other hand, we also have $\|q\|_{\mathcal{C}^{\alpha-1}([0,1]^d)} \leq M$. Deduce that

$$\sup_{n\geq 1}\|g_n\|_{\mathcal{B}^{\alpha-1}_{\infty,\infty}([0,1]^d)} \lesssim \sup_{n\geq 1}\|g_n\|_{\mathcal{C}^{\alpha-1}([0,1]^d)} \lesssim 1,$$

where the first inequality again uses Lemma 102, and the second inequality follows from Lemma 95. Now, let $\alpha_{n,\xi} = \int \xi g_n$ for all $\xi \in \Psi$. By following the same argument as in the first part of this proof, and using again the fact that $\|\psi_0\|_{\mathcal{B}^{\alpha+1}_{\infty,\infty}([0,1]^d)} \lesssim \lambda$, we may deduce that

$$\begin{aligned}(I) &\leq \sum_{j=J_n+1}^{\infty}\sum_{\xi\in\Psi_j}|\gamma_\xi \alpha_{n,\xi}| \\ &\leq \|\psi_0\|_{\mathcal{B}^{\alpha+1}_{\infty,\infty}([0,1]^d)}\|g_n\|_{\mathcal{B}^{\alpha-1}_{\infty,\infty}([0,1]^d)}\sum_{j=J_n+1}^{\infty} 2^{dj}2^{-j[(\alpha+1)+\frac{d}{2}]}2^{-j[(\alpha-1)+\frac{d}{2}]} \lesssim 2^{-2J_n\alpha}.\end{aligned}$$

Likewise, we have

$$(II) = \left|\int(\psi_{J_n} - \psi_0)q\right| \lesssim 2^{-2J_n\alpha},$$

and the claim follows from here. □

5.E Proofs of Two-Sample Stability Bounds

5.E.1 Proof of Proposition 23

Due to the absolute continuity of P and Q, the optimal transport map from Q to P is given by $S_0 = \nabla\varphi_0^*$. Furthermore, by absolute continuity of P, there exists an optimal transport map $\widehat{\sigma}$ from P to $\widehat{P}$. We clearly have,

$$(\widehat{\sigma}\circ S_0)_{\#}Q = \widehat{P}.$$

Also let $\widehat{\pi} \in \Pi(Q, \widehat{Q})$ be the optimal coupling between Q and $\widehat{Q}$, so that

$$(\widehat{\sigma} \circ S_0, Id)_{\#}\widehat{\pi} \in \Pi(\widehat{P}, \widehat{Q}).$$

We deduce,

$$\begin{aligned}
W_2^2(\widehat{P}, \widehat{Q}) &\leq \int \|\widehat{\sigma} \circ S_0(y) - z\|^2 \, d\widehat{\pi}(y,z) \\
&= \int \Big[\|\widehat{\sigma} \circ S_0(y) - S_0(y)\|^2 + \|S_0(y) - z\|^2 \Big] d\widehat{\pi}(y,z) \\
&\quad + 2 \int \langle \widehat{\sigma} \circ S_0(y) - S_0(y), S_0(y) - z \rangle d\widehat{\pi}(y,z). \qquad (5.37)
\end{aligned}$$

Notice that

$$\int \|\widehat{\sigma} \circ S_0(y) - S_0(y)\|^2 \, d\widehat{\pi}(y,z) = \int \|\widehat{\sigma}(x) - x\|^2 dP(x) = W_2^2(\widehat{P}, P). \qquad (5.38)$$

Furthermore, we have

$$\int \|S_0(y) - z\|^2 d\widehat{\pi}(y,z) \leq W_2^2(P,Q) + \int \psi_0 d(\widehat{Q} - Q) + \lambda W_2^2(\widehat{Q}, Q), \qquad (5.39)$$

by Lemma 40. Additionally, the cross term in equation (5.37) is bounded as follows.

Lemma 42. *We have,*

$$\begin{aligned}
&2 \int \langle \widehat{\sigma} \circ S_0(y) - S_0(y), S_0(y) - z \rangle d\widehat{\pi}(y,z) \\
&\qquad \leq \int \phi_0 d(\widehat{P} - P) + 2W_2(\widehat{P}, P) W_2(\widehat{Q}, Q) + (\lambda - 1) W_2^2(\widehat{P}, P).
\end{aligned}$$

We prove Lemma 42 in Appendix 5.E.2 below. By equations (5.37–5.39) and Lemma 42, we obtain

$$\begin{aligned}
W_2^2(\widehat{P}, \widehat{Q}) &\leq W_2^2(P,Q) + \lambda W_2^2(\widehat{P}, P) + \lambda W_2^2(\widehat{Q}, Q) \\
&\qquad + \int \psi_0 d(\widehat{Q} - Q) + \int \phi_0 d(\widehat{P} - P) + 2W_2(\widehat{P}, P) W_2(\widehat{Q}, Q) \\
&\leq W_2^2(P,Q) + \lambda \Big[W_2(\widehat{P}, P) + W_2(\widehat{Q}, Q) \Big]^2 + \int \psi_0 d(\widehat{Q} - Q) + \int \phi_0 d(\widehat{P} - P).
\end{aligned}$$

This proves the upper bound of the claim. To prove the lower bound, notice that, by the Kantorovich duality,

$$\begin{aligned}
W_2^2(\widehat{P}, \widehat{Q}) &\geq \int \phi_0 d\widehat{P} + \int \psi_0 d\widehat{Q} \\
&= \int \phi_0 dP + \int \psi_0 dQ + \int \phi_0 d(\widehat{P} - P) + \int \psi_0 d(\widehat{Q} - Q) \\
&= W_2^2(P,Q) + \int \phi_0 d(\widehat{P} - P) + \int \psi_0 d(\widehat{Q} - Q).
\end{aligned}$$

The claim follows. □

5.E.2 Proof of Lemma 42

Write

$$2\int \langle \widehat{\sigma} \circ S_0(y) - S_0(y), S_0(y) - z\rangle d\widehat{\pi}(y,z) = (I) + (II) + (III), \tag{5.40}$$

where

$$\begin{aligned}(I) &= 2\int \langle \widehat{\sigma} \circ S_0(y) - S_0(y), y - z\rangle d\widehat{\pi}(y,z)\\ (II) &= 2\int \langle \widehat{\sigma} \circ S_0(y) - S_0(y), -y\rangle d\widehat{\pi}(y,z)\\ (III) &= 2\int \langle \widehat{\sigma} \circ S_0(y) - S_0(y), S_0(y)\rangle d\widehat{\pi}(y,z).\end{aligned}$$

Regarding (I), the Cauchy-Schwarz inequality implies

$$\begin{aligned}(I) &\leq 2\left(\int \|\widehat{\sigma} \circ S_0(y) - S_0(y)\|^2 \, d\widehat{\pi}(y,z)\right)^{\frac{1}{2}} \left(\int \|y - z\|^2 \, d\widehat{\pi}(y,z)\right)^{\frac{1}{2}}\\ &= 2\left(\int \|\widehat{\sigma}(x) - x\|^2 \, dP(x)\right)^{\frac{1}{2}} \left(\int \|y - z\|^2 \, d\widehat{\pi}(y,z)\right)^{\frac{1}{2}}\\ &= 2W_2(\widehat{P}, P)W_2(\widehat{Q}, Q).\end{aligned} \tag{5.41}$$

Regarding term (II), recall that φ_0 satisfies assumption **A1(λ)**, thus we have

$$\frac{1}{2\lambda}\|x - y\|^2 \leq \varphi_0(y) - \varphi_0(x) - \big\langle T_0(x), y - x\big\rangle \leq \frac{\lambda}{2}\|x - y\|^2, \quad x, y \in \Omega,$$

We deduce that,

$$\begin{aligned}(II) &= 2\int \langle \widehat{\sigma} \circ S_0(y) - S_0(y), -T_0 \circ S_0(y)\rangle d\widehat{\pi}(y,z)\\ &\leq 2\int \left[\varphi_0(S_0(y)) - \varphi_0(\widehat{\sigma} \circ S_0(y)) + \frac{\lambda}{2}\|S_0(y) - \widehat{\sigma} \circ S_0(y)\|^2\right] d\widehat{\pi}(y,z)\\ &= \int 2\varphi_0 d(P - \widehat{P}) + \lambda W_2^2(\widehat{P}, P).\end{aligned} \tag{5.42}$$

Finally, term (III) satisfies

$$\begin{aligned}(III) &= \int \left[\|\widehat{\sigma} \circ S_0(y)\|^2 - \|\widehat{\sigma} \circ S_0(y) - S_0(y)\|^2 - \|S_0(y)\|^2\right] d\widehat{\pi}(y,z)\\ &= \int \|\cdot\|^2 \, d(\widehat{P} - P) - W_2^2(\widehat{P}, P).\end{aligned} \tag{5.43}$$

Combine equations (5.41)–(5.43) with equation (5.40) to deduce the claim. □

5.E.3 Proof of Proposition 24

Once again, denote by $S_0 = \nabla \varphi_0^*$ the optimal transport map from Q to P. Recall from the proof of Theorem 18 (equation (5.30)) that, due to assumption **A1(λ)**,

$$\frac{1}{2\lambda}\left\|x-y\right\|^2 \le \varphi_0^*(y) - \varphi_0^*(x) - \big\langle S_0(x), y-x\big\rangle \le \frac{\lambda}{2}\left\|x-y\right\|^2,$$

for all $x, y \in \Omega$. Now, we have,

$$\begin{aligned} W_2^2(P_n, Q_m) &= \sum_{i=1}^n \sum_{j=1}^m \widehat{\pi}_{ij} \|X_i - Y_j\|^2 \\ &= \sum_{i=1}^n \sum_{j=1}^m \widehat{\pi}_{ij} \Bigg[\|T_0(X_i) - X_i\|^2 \\ &\qquad\qquad + 2\langle T_0(X_i) - X_i, Y_j - T_0(X_i)\rangle + \|Y_j - T_0(X_i)\|^2 \Bigg]. \end{aligned}$$

Notice that

$$\begin{aligned} \mathbb{E}\left[\sum_{i=1}^n \sum_{j=1}^m \widehat{\pi}_{ij} \|T_0(X_i) - X_i\|^2\right] &= \mathbb{E}\left[\sum_{i=1}^n \left(\sum_{j=1}^m \widehat{\pi}_{ij}\right) \|T_0(X_i) - X_i\|^2\right] \\ &= \mathbb{E}\left[\frac{1}{n}\sum_{i=1}^n \|T_0(X_i) - X_i\|^2\right] = W_2^2(P, Q), \end{aligned}$$

where we have used the marginal constraint on the coupling $\widehat{\pi}$ in the first equality of the above display. Recalling that $\Delta_{nm} = \sum_{i=1}^n \sum_{j=1}^m \widehat{\pi}_{ij} \|X_i - Y_j\|^2$, thus we obtain,

$$\begin{aligned} &\mathbb{E}\Big[W_2^2(P_n,Q_m) - W_2^2(P, Q)\Big] \\ &\qquad = \mathbb{E}[\Delta_{nm}] + 2\mathbb{E}\left[\sum_{i=1}^n \sum_{j=1}^m \widehat{\pi}_{ij} \langle T_0(X_i) - X_i, Y_j - T_0(X_i)\rangle\right] \\ &\qquad = \mathbb{E}[\Delta_{nm}] + 2\mathbb{E}\left[\sum_{i=1}^n \sum_{j=1}^m \widehat{\pi}_{ij} \langle T_0(X_i) - S_0(T_0(X_i)), Y_j - T_0(X_i)\rangle\right]. \end{aligned}$$

Now,

$$2\langle -S_0(T_0(X_i)), Y_j - T_0(X_i)\rangle \ge 2\varphi_0^*(T_0(X_i)) - 2\varphi_0^*(Y_j) + \frac{1}{\lambda}\left\|T_0(X_i) - Y_j\right\|^2, \tag{5.44}$$

whence, we obtain,

$$\begin{aligned} \mathbb{E}\Big[W_2^2(P_n, Q_m) - W_2^2(P, Q)\Big] \ge \mathbb{E}[\Delta_{nm}] + \mathbb{E}\Bigg[\sum_{i=1}^n \sum_{j=1}^m \widehat{\pi}_{ij} \Big(& 2\varphi_0^*(T_0(X_i)) - 2\varphi_0^*(Y_j) \\ & + \frac{1}{\lambda}\|T_0(X_i) - Y_j\|^2 + 2\langle T_0(X_i), Y_j - T_0(X_i)\rangle \Big)\Bigg] \end{aligned}$$

Now, notice that

$$2\langle T_0(X_i), Y_j - T_0(X_i)\rangle = -\left\|T_0(X_i) - Y_j\right\|^2 + \left\|Y_j\right\|^2 - \left\|T_0(X_i)\right\|^2.$$

Thus, continuing from before, we have

$$\begin{aligned}
&\mathbb{E}\Big[W_2^2(P_n, Q_m) - W_2^2(P,Q)\Big] \\
&\qquad \geq \frac{1}{\lambda}\mathbb{E}[\Delta_{nm}] + \mathbb{E}\left[\sum_{i=1}^n \sum_{j=1}^m \widehat{\pi}_{ij}\Big(2\varphi_0^*(T_0(X_i)) - 2\varphi_0^*(Y_j) + \|Y_j\|^2 - \|T_0(X_i)\|^2\Big)\right] \\
&\qquad = \frac{1}{\lambda}\mathbb{E}[\Delta_{nm}] + \mathbb{E}\left[\frac{1}{m}\sum_{j=1}^m \Big(\, \|Y_j\|^2 - 2\varphi_0^*(Y_j)\Big)\right] \\
&\qquad\qquad\qquad - \mathbb{E}\left[\frac{1}{n}\sum_{i=1}^n \Big(\, \|T_0(X_i)\|^2 - 2\varphi_0^*(T_0(X_i))\Big)\right] = \frac{1}{\lambda}\mathbb{E}[\Delta_{nm}].
\end{aligned}$$

This proves one of the inequalities of the claim. To obtain the other, return to equation (5.44) and notice that one also has

$$2\langle -S_0(T_0(X_i)), Y_j - T_0(X_i)\rangle \leq 2\varphi_0^*(T_0(X_i)) - 2\varphi_0^*(Y_j) + \lambda \left\|T_0(X_i) - Y_j\right\|^2.$$

The proof then proceeds analogously. This proves that

$$\mathbb{E}[\Delta_{nm}] \asymp_\lambda \mathbb{E}\Big[W_2^2(P_n, Q_m) - W_2^2(P,Q)\Big].$$

To conclude, apply Proposition 23 to deduce

$$\begin{aligned}
&\mathbb{E}\Big[W_2^2(P_n, Q_m) - W_2^2(P,Q)\Big] \\
&\qquad \leq \mathbb{E}\int \phi_0 d(P_n - P) + \mathbb{E}\int \psi_0 d(Q_m - Q) + 2\lambda\Big[\mathbb{E}W_2^2(P_n, P) + \mathbb{E}W_2^2(Q_m, Q)\Big] \\
&\qquad = 2\lambda\Big[\mathbb{E}W_2^2(P_n, P) + \mathbb{E}W_2^2(Q_m, Q)\Big].
\end{aligned}$$

The above display is of the order $\kappa_{n\wedge m}$ due to equation (5.12). When we additionally assume that $\Omega = [0,1]^d$ and $\gamma^{-1} \leq p, q \leq \gamma$, we may instead bound it from above by $\overline{\kappa}_{n\wedge m}$, due to Corollary 19. The claim follows. □

5.E.4 Proof of Proposition 27

The claim follows along the same lines as the proofs of Theorem 18 and Proposition 23, thus we only provide a brief proof of the analogue of Theorem 18 over the torus. It will suffice to prove

$$\frac{1}{\lambda}\|\widehat{T} - T_0\|_{L^2(P)}^2 \leq \mathcal{W}_2^2(P, \widehat{Q}) - \mathcal{W}_2^2(P, Q) - \int \psi_0 d(\widehat{Q} - Q) \leq \lambda \mathcal{W}_2^2(\widehat{Q}, Q). \tag{5.45}$$

Recall that $\widehat{T}$ is the optimal transport map from P to $\widehat{Q}$. By Proposition 3(iii), we therefore have P-almost surely

$$d_{\mathbb{T}^d}(\widehat{T}(x), x) = \|\widehat{T}(x) - x\|, \quad d_{\mathbb{T}^d}(T_0(x), x) = \|T_0(x) - x\|, \quad x \in \mathbb{T}^d.$$

It follows that

$$\mathcal{W}_2^2(P, \widehat{Q}) - \mathcal{W}_2^2(P, Q) = \int \|\widehat{T}(x) - x\|^2 dP(x) - \int \|T_0(x) - x\|^2 dP(x).$$

From here, it follows identically as in the proof of Theorem 18 that

$$\mathcal{W}_2^2(P, \widehat{Q}) - \mathcal{W}_2^2(P, Q) \geq \frac{1}{\lambda}\|\widehat{T} - T_0\|_{L^2(P)}^2 + \int \psi_0 d(\widehat{Q} - Q).$$

To prove the second inequality in equation (5.45), let $\widehat{\pi}$ denote an optimal coupling between Q and $\widehat{Q}$ with respect to the cost $d_{\mathbb{T}^d}^2$. Notice similarly as before that Proposition 3(iii) implies

$$\mathcal{W}_2^2(P, Q) = \int \|S_0(y) - y\|^2 dQ(y), \quad \mathcal{W}_2^2(Q, \widehat{Q}) = \int \|y - z\|^2 d\widehat{\pi}(y, z),$$

thus, since $(S_0, Id)_{\#}\widehat{\pi} \in \Pi(P, \widehat{Q})$, and using the fact that $d_{\mathbb{T}^d} \leq \|\cdot\|$, we have

$$\begin{aligned}
&\mathcal{W}_2^2(P, \widehat{Q}) \\
&\quad \leq \int d_{\mathbb{T}^d}^2(S_0(y), z) d\widehat{\pi}(y, z) \\
&\quad \leq \int \|S_0(y) - z\|^2 d\widehat{\pi}(y, z) \\
&\quad = \int \|S_0(y) - y\|^2 dQ(y) + \int \|y - z\|^2 d\widehat{\pi}(y, z) + 2\int \big\langle S_0(y) - y, y - z\big\rangle d\widehat{\pi}(y, z) \\
&\quad = \mathcal{W}_2^2(P, Q) + \mathcal{W}_2^2(\widehat{Q}, Q) + 2\int \big\langle S_0(y) - y, y - z\big\rangle d\widehat{\pi}(y, z).
\end{aligned}$$

By the same argument as in Theorem 18, the cross term is bounded above by $(\lambda - 1)\mathcal{W}_2^2(\widehat{Q}, Q) + \int \psi_0 d(\widehat{Q} - Q)$, thus the claim follows. □

5.F Proofs of Upper Bounds for Two-Sample Empirical Estimators

In this Appendix, we prove Propositions 25 and 26. We begin with the following result.

Lemma 43. *Let Ω satisfy conditions **(S1)–(S2)**. Let $P \in \mathcal{P}_{\mathrm{ac}}(\Omega)$ admit a density p such that $\gamma^{-1} \leq p \leq \gamma$ for some $\gamma > 0$. Let $V_1, \dots, V_n$ denote the Voronoi partition in equation* (5.23)*, based on an i.i.d. sample $X_1, \dots, X_n \sim P$. Then, there exist constants $C_1, C_2 > 0$ depending only on $d, \gamma, \epsilon_0, \delta_0$ such that the following assertions hold.*

(i) For all $\delta \in (0,1)$*, we have,*

$$\mathbb{P}\left(\max_{1\leq i\leq n} P(V_i) \geq \frac{C_1}{n}\Big[d\log n + \log(1/\delta)\Big]\right) \leq \delta.$$

(ii) We have,

$$\mathbb{E}\left[\max_{1\leq i\leq n} \mathrm{diam}(V_i)^2\right] \leq C_2\left(\frac{\log n}{n}\right)^{\frac{2}{d}}.$$

Proof of Lemma 43. We shall make use of the relative Vapnik-Chervonenkis inequality (Vapnik, 2013; Bousquet, Boucheron, and Lugosi, 2003), in the following form stated by Chaudhuri and Dasgupta (2010).

Lemma 44. *Let* $\mathcal{B}$ *denote the set of balls in* $\mathbb{R}^d$*. Then, there exists a universal constant* $C > 0$ *such that for every* $\delta \in (0,1)$*, we have with probability at least* $1-\delta$ *that for all* $B \in \mathcal{B}$*,*

$$P(B) \geq \frac{C}{n}\left[d\log n + \log\left(\frac{1}{\delta}\right)\right] \implies P_n(B) > 0.$$

We now turn to the proof. Recall that Ω is a standard set by condition **(S2)**, and recall the constants $\epsilon_0, \delta_0 > 0$ therein. For any $1 \leq i \leq n$ and $x \in V_i \setminus \{X_i\}$, let $\rho_i(x) = (\epsilon_0/2d)\,\|x - X_i\|$. Since $\mathrm{diam}(\Omega) \leq \sqrt{d}$ by condition **(S1)**, we have $\rho_i(x) \leq \epsilon_0$. We also have $\rho_i(x) < \|x - X_i\|$, thus the balls $B(x, \rho_i(x))$ of radius $\rho_i(x)$ centered at x contain no sample points. Therefore, by Lemma 44, we have that for any $\delta \in (0,1)$, with probability at least $1-\delta$,

$$\max_{1\leq i\leq n}\sup_{x\in V_i} P\big(B(x_i, \rho_i(x))\big) \leq \frac{C}{n}\left[d\log n + \log\left(\frac{1}{\delta}\right)\right]. \tag{5.46}$$

Now, since $\gamma^{-1} \leq p \leq \gamma$, the assumption of standardness on Ω leads to the bound

$$P(B(x,\rho_i(x))) \geq \gamma^{-1}\mathcal{L}(B(x,\rho_i(x)) \cap \Omega) \geq \delta_0\gamma^{-1}\mathcal{L}(B(x,\rho_i(x))) \asymp \rho_i^d(x),$$

thus equation (5.46) reduces to

$$\max_{1\leq i\leq n}\sup_{x\in V_i} \rho_i^d(x) \leq \frac{C}{n}\left[d\log n + \log\left(\frac{1}{\delta}\right)\right].$$

Deduce from here that with probability at least $1-\delta$, for any $1 \leq i \leq n$ and $x, y \in V_i$,

$$\|x - y\| \leq \|x - X_i\| + \|y - Y_i\| \lesssim \rho_i(x) + \rho_i(y) \lesssim \left[\frac{d\log n + \log(1/\delta)}{n}\right]^{\frac{1}{d}}.$$

It follows that for some $C_1 > 0$ not depending on δ, we have with probability at least $1-\delta$,

$$\max_{1\leq i\leq n} \mathrm{diam}(V_i) \leq C_1\left[\frac{d\log n + \log(1/\delta)}{n}\right]^{\frac{1}{d}}.$$

To prove claim (i), notice that since the density of P is bounded from above, we also have with probability at least $1-\delta$,

$$\max_{1\leq i\leq n} P(V_i) \leq \gamma \max_{1\leq i\leq n} \mathcal{L}(V_i) \lesssim \max_{1\leq i\leq n} \mathrm{diam}(V_i)^d \lesssim \frac{1}{n}\Big[d\log n + \log(1/\delta)\Big].$$

To prove claim (ii), let $t_n = (2C_1^d(d+2)\log n/n)^{2/d}$. Set $\delta = n^d \exp\left(-\frac{u^d n}{C_1^d}\right)$ for any $u > 0$ to obtain

$$\begin{aligned}
\mathbb{E}\left[\max_{1\leq i\leq n} \mathrm{diam}(V_i)^2\right] &= \int_0^\infty \mathbb{P}\left(\max_{1\leq i\leq n} \mathrm{diam}(V_i)^2 \geq u\right) du \\
&\leq t_n + n^d \int_{t_n}^\infty \exp\left(-\frac{u^{\frac{d}{2}} n}{C_1^d}\right) du \\
&= t_n + \frac{4n^d}{d} \int_{t_n^{d/4}}^\infty \exp\left(-\frac{v^2 n}{C_1^d}\right) v^{\frac{4}{d}-1} dv \\
&\lesssim t_n + n^d \int_{t_n^{d/4}}^\infty \exp\left(-\frac{v^2 n}{2C_1^d}\right) dv \\
&\lesssim t_n + \frac{n^d}{\sqrt{n}} \exp\left(-\frac{t_n^{d/2} n}{2C_1^d}\right) \lesssim \left(\frac{\log n}{n}\right)^{\frac{2}{d}}.
\end{aligned}$$

The claim follows. □

5.F.1 Proof of Proposition 25

Abbreviate $\widehat{T}_{nm}^{\mathrm{1NN}}$ by $\widehat{T}_{nm}$. We have,

$$\begin{aligned}
\big\|\widehat{T}_{nm} - T_0\big\|_{L^2(P)}^2 &= \sum_{i=1}^n \int_{V_i} \|\widehat{T}_{nm}(x) - T_0(X_i) + T_0(X_i) - T_0(x)\big\|^2 dP(x) \\
&\lesssim \sum_{i=1}^n \int_{V_i} \Big[\|\widehat{T}_{nm}(x) - T_0(X_i)\|^2 + \|T_0(X_i) - T_0(x)\|^2\Big] dP(x).
\end{aligned}$$

To bound the first term, notice that,

$$\begin{aligned}
\sum_{i=1}^n \int_{V_i} \|\widehat{T}_{nm}(x) - T_0(X_i)\|^2 dP(x) &= \sum_{i=1}^n \int_{V_i} \left\|\sum_{j=1}^m (n\widehat{\pi}_{ij}) Y_j - T_0(X_i)\right\|^2 dP(x) \\
&= \sum_{i=1}^n P(V_i) \left\|\sum_{j=1}^m (n\widehat{\pi}_{ij}) Y_j - T_0(X_i)\right\|^2 \\
&\leq \sum_{i=1}^n P(V_i) \sum_{j=1}^m (n\widehat{\pi}_{ij}) \|Y_j - T_0(X_i)\|^2,
\end{aligned}$$

by convexity of $\|\cdot\|^2$. Therefore, setting $M_n = \max_{1\leq i\leq n} P(V_i)$, we obtain

$$\big\|\widehat{T}_{nm} - T_0\big\|_{L^2(P)}^2 \leq n\Delta_{nm}\left(\max_{1\leq i\leq n} P(V_i)\right) + \sum_{i=1}^n \int_{V_i} \|T_0(X_i) - T_0(x)\|^2 dP(x).$$

Since T_0 is λ-Lipschitz by condition **A1(λ)**, the claim is now a consequence of the following simple Lemma, which we isolate for future reference.

Lemma 45. *Under the conditions of the first claim of Proposition 25, we have for any λ-Lipschitz map $F : \Omega \to \Omega$,*

$$\mathbb{E}\left[\sum_{i=1}^n \int_{V_i} \|F(X_i) - F(x)\|^2 dP(x)\right] \lesssim_{\lambda,\gamma} (\log n/n)^{2/d},$$
$$\mathbb{E}\left[n\Delta_{nm}\left(\max_{1\leq i\leq n} P(V_i)\right)\right] \lesssim_{\lambda,\gamma,\epsilon_0,\delta_0} (\log n)\kappa_{n\wedge m}.$$

If we additionally assume that $\Omega = [0,1]^d$ and $\gamma^{-1} \leq q \leq \gamma$ over Ω, then

$$\mathbb{E}\left[n\Delta_{nm}\left(\max_{1\leq i\leq n} P(V_i)\right)\right] \lesssim_{\lambda,\gamma,\epsilon_0,\delta_0} (\log n)\bar{\kappa}_{n\wedge m}.$$

5.F.1.1 Proof of Lemma 45

The first quantity is easily bounded as follows,

$$\begin{aligned}
\mathbb{E}\left[\sum_{i=1}^n \int_{V_i} \|F(X_i) - F(x)\|^2 dP(x)\right] &\leq \lambda^2 \mathbb{E}\left[\sum_{i=1}^n \int_{V_i} \|X_i - x\|^2 dP(x)\right] \\
&\leq \lambda^2 \mathbb{E}\left[\sum_{i=1}^n P(V_i)\,\mathrm{diam}(V_i)^2\right] \\
&\leq \lambda^2 \mathbb{E}\left[\max_{1\leq i\leq n} \mathrm{diam}(V_i)^2\right] \lesssim \left(\frac{\log n}{n}\right)^{\frac{2}{d}},
\end{aligned}$$

where the final inequality is due to Lemma 21(ii). To bound the second quantity, let $M_n = \max_{1\leq i\leq n} P(V_i)$. By Lemma 21(i) with $\delta = 1/n^2$, there is a large enough constant $c > 0$ such that if $m_n = c\log n/n$, then $\mathbb{P}(M_n \geq m_n) \leq 1/n^2$. We have,

$$\mathbb{E}\left[nM_n\Delta_{nm}\right] = \mathbb{E}\left[nM_n I(M_n \geq m_n)\Delta_{nm}\right] + \mathbb{E}\left[nM_n I(M_n < m_n)\Delta_{nm}\right].$$

Notice that Δ_{nm} is bounded above by $\mathrm{diam}(\Omega)^2$, and $0 \leq M_n \leq 1$, thus, by Proposition 24,

$$\mathbb{E}\left[nM_n\Delta_{nm}\right] \lesssim n\mathbb{P}(M_n \geq m_n) + m_n n\mathbb{E}\left[\Delta_{nm}\right] \lesssim \frac{1}{n} + (\log n)\mathbb{E}\left[\Delta_{nm}\right] \lesssim (\log n)\kappa_{n\wedge m},$$

as desired. The final claim follows analogously. □

5.F.2 Proof of Proposition 26

Abbreviate $\widehat{T}_{nm}^{\mathrm{LS}}$ by $\widehat{T}_{nm}$. Notice first that we have

$$\begin{aligned}
\big\|\widehat{T}_{nm} - T_0\big\|_{L^2(P_n)}^2 &= \frac{1}{n}\sum_{i=1}^n \big\|\widehat{T}_{nm}(X_i) - T_0(X_i)\big\|^2 \\
&= \sum_{i=1}^n \sum_{j=1}^m \widehat{\pi}_{ij}\big\|\widehat{T}_{nm}(X_i) - T_0(X_i)\big\|^2 \\
&\lesssim \sum_{i=1}^n \sum_{j=1}^m \widehat{\pi}_{ij}\big\|\widehat{T}_{nm}(X_i) - Y_j\big\|^2 + \sum_{i=1}^n \sum_{j=1}^m \widehat{\pi}_{ij}\big\|Y_j - T_0(X_i)\big\|^2 \\
&\leq 2\sum_{i=1}^n \sum_{j=1}^m \widehat{\pi}_{ij}\big\|Y_j - T_0(X_i)\big\|^2 = 2\Delta_{nm}, \qquad (5.47)
\end{aligned}$$

where the final inequality follows by definition of $\widehat{T}_{nm}$, since $\varphi_0 \in \mathcal{J}_\lambda$ under assumption **A1($\boldsymbol{\lambda}$)**. Therefore,

$$\begin{aligned}
\big\|\widehat{T}_{nm} - T_0\big\|_{L^2(P)}^2 = \sum_{i=1}^n \int_{V_i} \big\|\widehat{T}_{nm} - T_0\big\|^2 dP \lesssim \sum_{i=1}^n \int_{V_i} \Big[\|\widehat{T}_{nm}(x) - \widehat{T}_{nm}(X_i)\|^2 \\
+ \|\widehat{T}_{nm}(X_i) - T_0(X_i)\|^2 + \|T_0(X_i) - T_0(x)\|^2\Big] dP(x).
\end{aligned}$$

By definition of $\mathcal{J}_\lambda$ and by assumption **A1($\boldsymbol{\lambda}$)**, $\widehat{T}_{nm}$ and T_0 are both λ-Lipschitz, thus by Lemma 45,

$$\begin{aligned}
\mathbb{E}\big\|\widehat{T}_{nm} - T_0\big\|_{L^2(P)}^2 &\lesssim \left(\frac{\log n}{n}\right)^{\frac{2}{d}} + \mathbb{E}\left[\sum_{i=1}^n \int_{V_i} \|\widehat{T}_{nm}(X_i) - T_0(X_i)\|^2 dP(x)\right] \\
&\leq \left(\frac{\log n}{n}\right)^{\frac{2}{d}} + \mathbb{E}\left[n\left(\max_{1\leq i\leq n} P(V_i)\right)\big\|\widehat{T}_{nm} - T_0\big\|_{L^2(P_n)}^2\right] \\
&\lesssim \left(\frac{\log n}{n}\right)^{\frac{2}{d}} + \mathbb{E}\left[n\left(\max_{1\leq i\leq n} P(V_i)\right)\Delta_{nm}\right],
\end{aligned}$$

where we used equation (5.47). Lemma 45 may now be applied to bound the right-hand term in the above display, leading to the claim. □

5.G Upper Bounds for Two-Sample Wavelet Estimators

In this section, we state and prove a result deferred from Section 5.3.3, regarding two-sample plugin estimators based on wavelet density estimation over the torus.

Unlike the boundary-corrected wavelet system used in Section 5.2.3, it will be convenient to introduce a simpler basis which guarantees that the density estimators are periodic. Recall that we described in Appendix A.3.2 how the standard Daubechies wavelet system may be periodized to obtain a set of $\mathbb{Z}^d$-periodic functions

$$\Psi^{\text{per}} = \{1\} \cup \bigcup_{j=0}^{\infty} \Psi_j^{\text{per}}, \quad \text{where} \quad \Psi_j^{\text{per}} = \big\{\xi_{jk\ell}^{\text{per}} : 0 \leq k \leq 2^{j-1}, \ell \in \{0,1\}^d \setminus \{0\}\big\},\ j \geq 0,$$

which forms an orthonormal basis of $L^2(\mathbb{T}^d)$ (Daubechies, 1992; Giné and Nickl, 2016). Whenever the densities p, q lie in $L^2(\mathbb{T}^d)$, they admit wavelet expansions of the form

$$p = 1 + \sum_{j=0}^{\infty} \sum_{\xi \in \Psi_j^{\text{per}}} \alpha_\xi \xi, \quad q = 1 + \sum_{j=0}^{\infty} \sum_{\xi \in \Psi_j^{\text{per}}} \beta_\xi \xi,$$

where $\alpha_\xi = \int \xi dP$ and $\beta_\xi = \int \xi dQ$. We then define the wavelet density estimators

$$\widetilde{p}_n^{(\text{per})} = 1 + \sum_{j=0}^{J_n} \sum_{\xi \in \Psi_j^{\text{per}}} \widehat{\alpha}_\xi \xi, \quad \widetilde{q}_m^{(\text{per})} = 1 + \sum_{j=0}^{J_m} \sum_{\xi \in \Psi_j^{\text{per}}} \widehat{\beta}_\xi \xi,$$

where $\widehat{\alpha}_\xi = \int \xi dP_n$ and $\widehat{\beta}_\xi = \int \xi dQ_m$. By orthonormality of Ψ^{per}, it is straightforward to see that $\widetilde{p}_n^{(\text{per})}, \widetilde{q}_m^{(\text{per})}$ integrate to unity, but may nevertheless be negative. We therefore define the final density estimators by

$$\widehat{p}_n^{(\text{per})} \propto \widetilde{p}_n^{(\text{per})} I(\widetilde{p}_n^{(\text{per})} \geq 0), \quad \widehat{q}_m^{(\text{per})} \propto \widetilde{q}_m^{(\text{per})} I(\widetilde{q}_m^{(\text{per})} \geq 0), \tag{5.48}$$

where the proportionality constants are to be chosen such that $\widehat{p}_n^{(\text{per})}$ and $\widehat{q}_m^{(\text{per})}$ are probability densities, which respectively induce probability distributions $\widehat{P}_n^{(\text{per})}, \widehat{Q}_m^{(\text{per})} \in \mathcal{P}_{\text{ac}}(\mathbb{T}^d)$. Once again, we drop all superscripts "per" whenever the choice of wavelet basis is unambiguous. We state the following bound for the two-sample estimator $\widehat{T}_{nm} \equiv \widehat{T}_{nm}^{(\text{per})}$ in equation (5.25), together with the associated plugin estimator of the squared Wasserstein distance. Recall the sequence $R_{T,n}(\alpha)$ defined in Theorem 19.

Theorem 22 (Two-Sample Wavelet Estimators)**.** Let $P, Q \in \mathcal{P}_{\text{ac}}(\mathbb{T}^d)$ admit densities $p, q \in \mathcal{C}^{\alpha-1}(\mathbb{T}^d; M, \gamma)$ for some $\alpha > 1$ and $M, \gamma > 0$. Assume $2^{J_n} \asymp n^{\frac{1}{d+2(\alpha-1)}}$. Then, there exists a constant $C > 0$ depending only on M, γ, α such that the following statements hold.

(i) (Optimal Transport Maps) We have,

$$\mathbb{E}\big\|\widehat{T}_{nm} - T_0\big\|_{L^2(P)}^2 \leq C R_{T,n\wedge m}(\alpha).$$

(ii) (Wasserstein Distances) When $\alpha \notin \mathbb{N}$, we have

$$\begin{aligned}\big|\mathbb{E}\mathcal{W}_2^2(\widehat{P}_n, \widehat{Q}_m) - \mathcal{W}_2^2(P,Q)\big| &\leq C R_{T,n\wedge m}(\alpha), \\ \mathbb{E}\big|\mathcal{W}_2^2(\widehat{P}_n, \widehat{Q}_m) - \mathcal{W}_2^2(P,Q)\big|^2 &\leq \left[C R_{T,n\wedge m}(\alpha) + \sqrt{\frac{\mathrm{Var}_P[\phi_0(X)]}{n} + \frac{\mathrm{Var}_Q[\psi_0(Y)]}{m}}\right]^2.\end{aligned}$$

The proof appears in Appendix 5.G.1. Theorem 22 shows that the plugin estimators $\widehat{T}_{nm}$ and $\mathcal{W}_2^2(\widehat{P}_n, \widehat{Q}_m)$ achieve analogous convergence rates as in the one-sample setting. Similarly as in Section 5.2.3, we may deduce from Theorem 22(ii) and Lemma 37 that

$$\mathbb{E}\big|\mathcal{W}_2^2(\widehat{P}_n, \widehat{Q}_m) - \mathcal{W}_2^2(P, Q)\big| \lesssim_{M,\gamma,\alpha} R_{T,n\wedge m}(\alpha) + (n \wedge m)^{-1/2}\mathcal{W}_2(P, Q).$$

Thus, in the high-smoothness regime $2(\alpha + 1) > d$, the risk of $\mathcal{W}_2^2(\widehat{P}_n, \widehat{Q}_m)$ decays at a rate which adapts to the magnitude of the Wasserstein distance between P and Q.

If one is willing to place assumptions on the regularity of the potentials φ_0 and φ_0^*, Theorem 22(ii) may be extended to the case where the sampling domain is taken to be the unit cube $[0, 1]^d$, as we show next. Such a result is made possible by the fact that Proposition 23 does not require any regularity of the fitted potentials. On the other hand, we do not know how to obtain an analogue of Theorem 22(i) over domains in $\mathbb{R}^d$. Let $P, Q \in \mathcal{P}_{\mathrm{ac}}([0, 1]^d)$, and denote by $\widehat{P}_n^{(\mathrm{bc})}$ and $\widehat{Q}_m^{(\mathrm{bc})}$ the boundary corrected wavelet estimators defined in Section 5.2.3.

Proposition 28. Let $P, Q \in \mathcal{P}_{\mathrm{ac}}([0, 1]^d)$ admit densities $p, q \in \mathcal{C}^{\alpha-1}([0, 1]^d; M, \gamma)$ for some $\alpha > 1$ and $M, \gamma > 0$. Assume further that for some $\lambda > 0$,

$$\varphi_0, \varphi_0^* \in \mathcal{C}^{\alpha+1}([0, 1]^d; \lambda). \tag{5.49}$$

Let $2^{J_n} \asymp n^{\frac{1}{d+2(\alpha-1)}}$. Then, there exists a constant $C > 0$ depending only on $M, \lambda, \gamma, \alpha$ such that,

$$\begin{aligned}\big|\mathbb{E}W_2^2(\widehat{P}_n^{(\mathrm{bc})}, \widehat{Q}_m^{(\mathrm{bc})}) - W_2^2(P, Q)\big| &\leq C R_{T,n\wedge m}(\alpha),\\ \mathbb{E}\big|W_2^2(\widehat{P}_n^{(\mathrm{bc})}, \widehat{Q}_m^{(\mathrm{bc})}) - W_2^2(P, Q)\big|^2 &\leq \left[C R_{T,n\wedge m}(\alpha) + \sqrt{\frac{\mathrm{Var}_P[\phi_0(X)]}{n} + \frac{\mathrm{Var}_Q[\psi_0(Y)]}{m}}\right]^2.\end{aligned}$$

The proof follows along similar lines as that of Theorem 22(ii), which will be given below, and is therefore omitted.

Condition (5.49) places a smoothness assumption on φ_0^* in addition to φ_0. If our analysis could be carried out over a domain $\Omega \subseteq \mathbb{R}^d$ with smooth boundary, then, under appropriate boundary conditions on the potentials and under the assumptions made on p, q, standard Schauder theory (Gilbarg and Trudinger, 2001) could be applied to the Monge-Ampère equation to obtain that $\varphi_0^* \in \mathcal{C}^{\alpha+1}(\Omega)$ as soon as $\varphi_0 \in \mathcal{C}^2(\Omega)$, with uniform Hölder norms (see Proposition 9.1 of Caffarelli and Cabré (1995)). We do not know whether analogues of such results can be applied over the hypercube $[0, 1]^d$, thus we have placed assumptions both on φ_0 and its convex conjugate.

We now turn to the proof of Theorem 22. We first note that the one-sample results from Section 5.2.3 may readily be extended to the optimal transport problem over $\mathbb{T}^d$.

Proposition 29. Assume $P, Q \in \mathcal{P}_{\mathrm{ac}}(\mathbb{T}^d)$ admit densities $p, q \in \mathcal{C}^{\alpha-1}(\mathbb{T}^d; M, \gamma)$ for some $\alpha > 1$, $\alpha \notin \mathbb{N}$, and $M, \gamma > 0$. Let $\widehat{q}_m = \widehat{q}_m^{(\mathrm{per})}$ be the periodic wavelet estimator defined in equation (5.48), and let $\widehat{Q}_m$ be the induced probability distribution. Let

$$\overline{T}_m = \operatorname*{argmin}_{T \in \mathcal{T}(P, \widehat{Q}_m)} \int d^2_{\mathbb{T}^d}(x, T(x)) dP(x).$$

Furthermore, let $2^{J_m} \asymp m^{\frac{1}{2(\alpha-1)+d}}$. Then, there exists a constant $C > 0$ depending only on M, γ, α such that the following statements hold.

(i) We have $\mathbb{E}\mathcal{W}_2^2(\widehat{Q}_m, Q) \leq CR_{T,m}(\alpha)$ and $\mathbb{E}\mathcal{W}_2^4(\widehat{Q}_m, Q) \leq CR^2_{T,m}(\alpha)$.

(ii) We have,

$$\left| \mathbb{E} \int \psi_0 d(\widehat{Q}_m - Q) \right| \leq C2^{-2J_m \alpha}$$
$$\left| \mathrm{Var}\left[\int \psi_0 d(\widehat{Q}_m - Q) \right] - \frac{\mathrm{Var}_Q[\psi_0(Y)]}{m} \right| \leq \frac{C2^{-2J_m\alpha}}{m}.$$

(iii) We have,

$$\begin{aligned} \mathbb{E}\|\overline{T}_m - T_0\|^2_{L^2(P)} &\leq CR_{T,m}(\alpha), \\ \left|\mathbb{E}\mathcal{W}_2^2(P, \widehat{Q}_m) - \mathcal{W}_2^2(P,Q)\right| &\leq CR_{T,m}(\alpha), \\ \mathbb{E}\left|\mathcal{W}_2^2(P, \widehat{Q}_m) - \mathcal{W}_2^2(P,Q)\right|^2 &\leq \left[CR_{T,m}(\alpha) + \sqrt{\frac{\mathrm{Var}_Q[\psi_0(Y)]}{m}} \right]^2. \end{aligned}$$

Notice that the only properties of the boundary-correct wavelet basis used in the proofs of Lemma 106 and Theorem 19 are those contained in Lemmas 101 and Lemma 104 of Appendix A.3, which are also stated to hold for the periodic wavelet basis. The proof of Proposition 29 is therefore a direct extension of these results. Notice that, unlike Theorem 19, we no longer require any conditions on the smoothnes of φ_0 itself, due to the torus regularity result of Theorem 4. Indeed, under the assumptions of Proposition 29, the latter implies that there exists a constant $C' > 0$ depending only on α, γ, M such that $\|\varphi_0\|_{\mathcal{C}^{\alpha+1}(\mathbb{T}^d)} \leq C'$, assuming $\alpha \notin \mathbb{N}$.

5.G.1 Proof of Theorem 22

Throughout the proof, we use the abbreviations

$$F(\widehat{P}_n) = \int \phi_0 d(\widehat{P}_n - P), \quad L(\widehat{Q}_m) = \int \psi_0 d(\widehat{Q}_m - Q).$$

We begin by proving part (ii). Under the assumptions of this case, Theorem 4 implies that $\|\varphi_0\|_{\mathcal{C}^{\alpha+1}(\mathbb{T}^d)} \leq M_0$ for a universal constant $M_0 > 0$ depending only on α, γ and M. In particular, it also follows from Proposition 3(vii) that φ_0 is strongly convex, and thus satisfies condition **A1(λ)** for some $\lambda > 0$ depending only on M_0 and γ. We may therefore invoke the two-sample stability bound over $\mathbb{T}^d$ in Proposition 24 (arising from Proposition 23) to deduce

$$\begin{aligned} F(\widehat{P}_n) + L(\widehat{Q}_m) &\leq \mathcal{W}_2^2(\widehat{P}_n, \widehat{Q}_m) - \mathcal{W}_2^2(P, Q) \\ &\leq F(\widehat{P}_n) + L(\widehat{Q}_m) + 2\lambda \left[\mathcal{W}_2^2(\widehat{P}_n, P) + \mathcal{W}_2^2(\widehat{Q}_m, Q)\right]. \end{aligned}$$

From Proposition 22(ii), it can be deduced that

$$\left|\mathbb{E}F(\widehat{P}_n)\right| \vee \left|\mathbb{E}L(\widehat{Q}_m)\right| \lesssim R_{T,n\wedge m}(\alpha) \tag{5.50}$$

$$\mathrm{Var}\left[F(\widehat{P}_n)\right] \leq \frac{\mathrm{Var}_P[\phi_0(X)]}{n} + CR_{T,n}^2(\alpha) \tag{5.51}$$

$$\mathrm{Var}\left[L(\widehat{Q}_m)\right] \leq \frac{\mathrm{Var}_Q[\psi_0(Y)]}{m} + CR_{T,m}^2(\alpha), \tag{5.52}$$

for a constant $C > 0$ depending only on M, γ, α, whose value we allow to change from line to line in the remainder of the proof. Thus, recalling Proposition 22(i),

$$\begin{aligned} &\left|\mathbb{E}\mathcal{W}_2^2(\widehat{P}_n,\widehat{Q}_m) - \mathcal{W}_2^2(P, Q)\right| \\ &\qquad \lesssim \left|\mathbb{E}F(\widehat{P}_n)\right| + \left|\mathbb{E}L(\widehat{Q}_m)\right| + \mathbb{E}\mathcal{W}_2^2(\widehat{P}_n, P) + \mathbb{E}\mathcal{W}_2^2(\widehat{Q}_m, Q) \lesssim R_{T,n\wedge m}(\alpha). \end{aligned}$$

Furthermore,

$$\begin{aligned} &\mathbb{E}\left|\mathcal{W}_2^2(\widehat{P}_n, \widehat{Q}_m) - \mathcal{W}_2^2(P, Q)\right|^2 \\ &\quad \leq \mathbb{E}\left[\left(\left|F(\widehat{P}_n)\right| + \left|L(\widehat{Q}_m)\right| + 2\lambda\big(\mathcal{W}_2^2(\widehat{P}_n, P) + \mathcal{W}_2^2(\widehat{Q}_m, Q)\big)\right)^2\right] =: (I) + (II) + (III), \end{aligned}$$

where

$$\begin{aligned} (I) &= \mathbb{E}\left[\left(\left|F(\widehat{P}_n)\right| + \left|L(\widehat{Q}_m)\right|\right)^2\right] \\ (II) &= 4\lambda^2\mathbb{E}\left[\left(\mathcal{W}_2^2(\widehat{P}_n, P) + \mathcal{W}_2^2(\widehat{Q}_m, Q)\right)^2\right] \\ (III) &= 4\lambda\mathbb{E}\left[\left(\mathcal{W}_2^2(\widehat{P}_n, P) + \mathcal{W}_2^2(\widehat{Q}_m, Q)\right)\left(\left|F(\widehat{P}_n)\right| + \left|L(\widehat{Q}_m)\right|\right)\right]. \end{aligned}$$

Regarding term (I), recall that we have assumed that X_i is independent of Y_j for all $i, j = 1, \dots, n$. Therefore, using equations (5.50–5.52),

$$\begin{aligned} (I) &= \mathbb{E}\big[F^2(\widehat{P}_n)\big] + \mathbb{E}\big[L^2(\widehat{Q}_m)\big] + 2\mathbb{E}\big|F(\widehat{P}_n)L(\widehat{Q}_m)\big| \\ &= \mathbb{E}\big[F^2(\widehat{P}_n)\big] + \mathbb{E}\big[L^2(\widehat{Q}_m)\big] + 2\mathbb{E}\big|F(\widehat{P}_n)\big|\mathbb{E}\big|L(\widehat{Q}_m)\big| \\ &= \mathrm{Var}\left[F(\widehat{P}_n)\right] + \mathrm{Var}\left[L(\widehat{Q}_m)\right] + \left|\mathbb{E}F(\widehat{P}_n)\right|^2 + \left|\mathbb{E}L(\widehat{Q}_m)\right|^2 + 2\mathbb{E}\big|F(\widehat{P}_n)\big|\mathbb{E}\big|L(\widehat{Q}_m)\big| \\ &\leq \frac{\mathrm{Var}_P[\phi_0(X)]}{n} + \frac{\mathrm{Var}_Q[\psi_0(Y)]}{m} + CR_{T,n\wedge m}^2(\alpha). \end{aligned}$$

Furthermore, by Proposition 22(i), we have

$$(II) \leq 8\lambda^2\left(\mathbb{E}\mathcal{W}_2^4(\widehat{P}_n, P) + \mathbb{E}\mathcal{W}_2^4(\widehat{Q}_m, Q)\right) \leq CR_{T,n\wedge m}^2(\alpha),$$

and, using the Cauchy-Schwarz inequality and equations (5.50–5.52), we obtain

$$\begin{aligned}
(III) &\le C\sqrt{\Big(\mathbb{E}\mathcal{W}_2^4(\widehat{P}_n, P) + \mathbb{E}\mathcal{W}_2^4(\widehat{Q}_m, Q)\Big)\Big(\mathbb{E}\big|F(\widehat{P}_n)\big|^2 + \mathbb{E}\big|L(\widehat{Q}_m)\big|^2\Big)} \\
&\le C\sqrt{R_{T,n\wedge m}^2(\alpha)\left(CR_{T,n\wedge m}^2(\alpha) + \frac{\mathrm{Var}_P[\phi_0(X)]}{n} + \frac{\mathrm{Var}_Q[\psi_0(Y)]}{m}\right)} \\
&\le CR_{T,n\wedge m}^2(\alpha) + CR_{T,n\wedge m}(\alpha)\sqrt{\frac{\mathrm{Var}_P[\phi_0(X)]}{n} + \frac{\mathrm{Var}_Q[\psi_0(Y)]}{m}}.
\end{aligned}$$

Deduce that

$$(I) + (II) + (III) \le \left(CR_{T,n\wedge m}(\alpha) + \sqrt{\frac{\mathrm{Var}_P[\phi_0(X)]}{n} + \frac{\mathrm{Var}_Q[\psi_0(Y)]}{m}}\right)^2.$$

Claim (ii) follows from here.

To prove part (i), we shall make use of the one-sample optimal transport problem from P to $\widehat{Q}_m$. Denote by $\bar{\varphi}_m$ an optimal Brenier potential for this problem, so that $\overline{T}_m = \nabla\bar{\varphi}_m$ is the optimal transport map pushing P forward onto $\widehat{Q}_m$, with respect to the cost function $d_{\mathbb{T}^d}^2$. Furthermore, denote by

$$\bar{\phi}_m = \|\cdot\|^2 - 2\bar{\varphi}_m, \quad \bar{\psi}_m = \|\cdot\|^2 - 2\bar{\varphi}_m^*,$$

a corresponding pair of optimal Kantorovich potentials. We proceed with three steps.

Step 1: Regularity of the Fitted Potentials. Recall that $\alpha > 1$, and fix $\epsilon \in (0, 1 \wedge \frac{\alpha-1}{2})$. By Lemma 104 and Lemma 105, under our choice of threshold J_n, and under the assumption $p, q \in \mathcal{C}^{\alpha-1}(\mathbb{T}^d; M, \gamma)$, it can be deduced that the event

$$\begin{aligned}
E_{nm} = \{\widetilde{p}_n = \widehat{p}_n\} &\cap \{\widetilde{q}_m = \widehat{q}_m\} \\
&\cap \Big\{\widetilde{p}_n, \widetilde{q}_m \ge 1/(2\gamma) \text{ over } \mathbb{T}^d\Big\} \\
&\cap \Big\{\|\widetilde{p}_n\|_{\mathcal{B}_{\infty,\infty}^{\epsilon}(\mathbb{T}^d)} \le 2\|p\|_{\mathcal{B}_{\infty,\infty}^{\alpha-1}(\mathbb{T}^d)}\Big\} \cap \Big\{\|\widetilde{q}_m\|_{\mathcal{B}_{\infty,\infty}^{\epsilon}(\mathbb{T}^d)} \le 2\|q\|_{\mathcal{B}_{\infty,\infty}^{\alpha-1}(\mathbb{T}^d)}\Big\}
\end{aligned}$$

satisfies $\mathbb{P}(E_{nm}^{\mathsf{c}}) \lesssim (n \wedge m)^{-2}$. Note that $\epsilon \notin \mathbb{N}$, thus by Lemma 102, we have on the event E_{nm},

$$\|\widehat{q}_m\|_{\mathcal{C}^{\epsilon}(\mathbb{T}^d)} \lesssim \|\widehat{q}_m\|_{\mathcal{B}_{\infty,\infty}^{\epsilon}(\mathbb{T}^d)} \lesssim \|q\|_{\mathcal{B}_{\infty,\infty}^{\alpha-1}(\mathbb{T}^d)} \lesssim \|q\|_{\mathcal{C}^{\alpha-1}(\mathbb{T}^d)} \le M,$$

and similarly for $\widehat{p}_n$. Thus, there exists $M_1 > 0$ depending only on M, γ such that

$$\|\widehat{p}_n\|_{\mathcal{C}^{\epsilon}(\mathbb{T}^d)}, \|\widehat{q}_m\|_{\mathcal{C}^{\epsilon}(\mathbb{T}^d)} \le M_1, \quad \text{on } E_{nm}.$$

Under the preceding display, together with the smoothness assumptions on the population densities p, q themselves, and the fact that $\widehat{p}_n, \widehat{q}_m, p, q \ge 1/(2\gamma)$ over $\mathbb{T}^d$ on the event E_{nm}, we may apply the regularity Theorem 4 to deduce that there exists a constant $M_2 > 0$ depending only on M_0, M_1, γ such that for all $n, m \ge 1$,

$$\|\varphi_0\|_{\mathcal{C}^{2+\epsilon}([0,1]^d)} \vee \|\widehat{\varphi}_{nm}\|_{\mathcal{C}^{2+\epsilon}([0,1]^d)} \vee \|\bar{\varphi}_m\|_{\mathcal{C}^{2+\epsilon}([0,1]^d)} \le M_2, \tag{5.53}$$

on E_{nm}. Deduce from Proposition 3(i) that the Hessians of the above potentials are uniformly bounded over $\mathbb{R}^d$. Further apply Proposition 3(vii) to deduce that $\widehat{\varphi}_{nm}$ and $\overline{\varphi}_m$ satisfy the curvature condition **A1($\boldsymbol{\lambda}$)** almost surely, up to modifying the value of $\lambda > 0$ in terms of M_2 and γ, namely:

$$\lambda^{-1} I_d \preceq \nabla^2 \varphi_0(x), \nabla^2 \overline{\varphi}_m(x), \nabla^2 \widehat{\varphi}_{nm}(x) \preceq \lambda I_d, \quad \text{for all } x \in \mathbb{R}^d;\ n, m \geq 1, \tag{5.54}$$

on the event E_{nm}.

Step 2: Reduction to Optimal Transport Problems with Same Source Distribution. In order to appeal to the one-sample stability bounds of Theorem 18, write

$$\big\|\widehat{T}_{nm} - T_0\big\|_{L^2(P)}^2 \lesssim \big\|\widehat{T}_{nm} - \overline{T}_m\big\|_{L^2(P)}^2 + \big\|\overline{T}_m - T_0\big\|_{L^2(P)}^2. \tag{5.55}$$

The first term in the above display compares transport maps which are optimal for distinct source distributions. We therefore proceed with the following reduction, over the event E_{nm}:

$$\begin{aligned}
\big\|\widehat{T}_{nm} - \overline{T}_m\big\|_{L^2(P)}^2 &= \int_{\mathbb{T}^d} \big\|\widehat{T}_{nm}(x) - \overline{T}_m(x)\big\|^2 dP(x) \\
&= \int_{\mathbb{T}^d} \big\|\widehat{T}_{nm}(\overline{T}_m^{-1}(y)) - y\big\|^2 d\widehat{Q}_m(y) \\
&= \int_{\mathbb{T}^d} \big\|\widehat{T}_{nm}(\overline{T}_m^{-1}(y)) - \widehat{T}_{nm}(\widehat{T}_{nm}^{-1}(y))\big\|^2 d\widehat{Q}_m(y),
\end{aligned} \tag{5.56}$$

where the second line follows from the fact that $(\overline{T}_m)_\# P = \widehat{Q}_m$, and the third follows by invertibility of $\widehat{T}_{nm}$, which is ensured by the strong convexity of $\widehat{\varphi}_{nm}$ in equation (5.54). This same equation implies that, on the event E_{nm}, $\widehat{T}_{nm} = \nabla \widehat{\varphi}_{nm}$ is Lipschitz with a uniform constant. It follows that

$$\big\|\widehat{T}_{nm} - \overline{T}_m\big\|_{L^2(P)}^2 \lesssim \int_{\mathbb{T}^d} \big\|\widehat{T}_{nm}^{-1}(y) - \overline{T}_m^{-1}(y)\big\|^2 d\widehat{Q}_m(y) = \big\|\widehat{T}_{nm}^{-1} - \overline{T}_m^{-1}\big\|_{L^2(\widehat{Q}_m)}^2. \tag{5.57}$$

Step 3: Stability Bounds. Due to the inequalities (5.54), the stability bounds of Proposition 27 (arising from Theorem 18) imply

$$\big\|\widehat{T}_{nm}^{-1} - \overline{T}_m^{-1}\big\|_{L^2(\widehat{Q}_m)}^2 \leq \lambda^2 \mathcal{W}_2^2(\widehat{P}_n, P), \quad \big\|\overline{T}_m - T_0\big\|_{L^2(P)}^2 \leq \lambda^2 \mathcal{W}_2^2(\widehat{Q}_m, Q). \tag{5.58}$$

Thus, combined with equations (5.55) and (5.57), we have on the event E_{nm},

$$\big\|\widehat{T}_{nm} - T_0\big\|_{L^2(P)}^2 \lesssim \mathcal{W}_2^2(\widehat{P}_n, P) + \mathcal{W}_2^2(\widehat{Q}_m, Q).$$

We deduce,

$$\begin{aligned}
\mathbb{E}\big\|\widehat{T}_{nm} - T_0\big\|_{L^2(P)}^2 &= \mathbb{E}\left[\big\|\widehat{T}_{nm} - T_0\big\|_{L^2(P)}^2 I_{E_{nm}}\right] + \mathbb{E}\left[\big\|\widehat{T}_{nm} - T_0\big\|_{L^2(P)}^2 I_{E_{nm}^{\mathsf{c}}}\right] \\
&\lesssim \mathbb{E}\left[\mathcal{W}_2^2(\widehat{P}_n, P) I_{E_{nm}}\right] + \mathbb{E}\left[\mathcal{W}_2^2(\widehat{Q}_m, Q) I_{E_{nm}}\right] + \mathbb{P}(E_{nm}^{\mathsf{c}}) \\
&\leq \mathbb{E}\left[\mathcal{W}_2^2(\widehat{P}_n, P)\right] + \mathbb{E}\left[\mathcal{W}_2^2(\widehat{Q}_m, Q)\right] + \mathbb{P}(E_{nm}^{\mathsf{c}}) \\
&\lesssim R_{T,n}(\alpha) + R_{T,m}(\alpha) + (n \wedge m)^{-2} \lesssim R_{T,n\wedge m}(\alpha),
\end{aligned}$$

where we made use of Proposition 22(i) on the final line. The claim follows. □

5.H Proofs of Upper Bounds for Two-Sample Kernel Estimators

The goal of this Appendix is to prove Theorem 20. For ease of notation, we omit the superscript "ker" in all kernel-based estimators, and write

$$p_{h_n}(x) = \mathbb{E}[\widetilde{p}_n(x)] = (p \star K_{h_n})(x), \quad q_{h_m}(y) = \mathbb{E}[\widetilde{q}_m(y)] = (q \star K_{h_m})(y), \quad x, y \in \mathbb{T}^d.$$

The following result was anticipated by Divol (2021), who derived a Fourier-analytic proof of the convergence rate of the empirical measure under the Wasserstein distance on $\mathbb{T}^d$. Our proof follows along similar lines, and is simplified by the fact that we work only with the Wasserstein distance of second order, but is complicated by the fact that we require general exponents $\rho \geq 0$.

Lemma 46. *Let $s > 0$. Assume $P \in \mathcal{P}_{\mathrm{ac}}(\mathbb{T}^d)$ admits a density p such that*

$$\|p\|_{H^s(\mathbb{T}^d)} \leq R < \infty, \qquad 0 < \gamma^{-1} \leq p \leq \gamma < \infty.$$

Assume further that the kernel K satisfies condition **K($s+1$)** *for some $\kappa > 0$. Set $h_n \asymp n^{\frac{1}{2s+d}}$. Then, for any $\rho \geq 0$,*

$$\mathbb{E}\mathcal{W}_2^\rho(\widehat{P}_n, P) \lesssim_{R,\rho,\gamma,s} \begin{cases} n^{-\frac{\rho(s+1)}{2s+d}}, & d \geq 3 \\ (\log n/n)^{\rho/2}, & d = 2 \\ (1/n)^{\rho/2}, & d = 1. \end{cases}$$

Proof. By Jensen's inequality, it suffices to prove the claim for $\rho \geq 2$. It is a direct consequence of Proposition 42 and the assumption $\gamma^{-1} \leq p \leq \gamma$ that the event $A_n = \{\widehat{p}_n = \widetilde{p}_n\}$ satisfies $\mathbb{P}(A_n) \lesssim 1/n^2$. Furthermore, recall from equation (5.9), arising from the work of Peyre (2018), that

$$\mathcal{W}_2(\widehat{P}_n, P) \lesssim \|\widehat{p}_n - p\|_{\dot{H}^{-1}}.$$

We therefore have,

$$\begin{aligned} \mathbb{E}\mathcal{W}_2^\rho(\widehat{P}_n, P) &= \mathbb{E}\Big[\mathcal{W}_2^\rho(\widehat{P}_n, P) I_{A_n}\Big] + \mathbb{E}\Big[\mathcal{W}_2^\rho(\widehat{P}_n, P) I_{A_n^c}\Big] \\ &\lesssim \mathbb{E}\Big[\|\widehat{p}_n - p\|_{\dot{H}^{-1}}^\rho I_{A_n}\Big] + 1/n^2 \\ &= \mathbb{E}\Big[\|\widetilde{p}_n - p\|_{\dot{H}^{-1}}^\rho I_{A_n}\Big] + 1/n^2 \\ &\lesssim \|p_{h_n} - p\|_{\dot{H}^{-1}}^\rho + \mathbb{E}\|\widetilde{p}_n - p_{h_n}\|_{\dot{H}^{-1}}^\rho + 1/n^2 \\ &\lesssim h_n^{\rho(s+1)} + \mathbb{E}\|\widetilde{p}_n - p_{h_n}\|_{\dot{H}^{-1}}^\rho + 1/n^2, \end{aligned} \tag{5.59}$$

where we used proposition 43 to bound the bias term in the final line, together with the assumption **K($s+1$)**. To bound the variance term, write $\mathbb{E}\|\widetilde{p}_n - p_{h_n}\|_{\dot{H}^{-1}}^\rho \lesssim S_{n,1} + S_{n,2}$, where

$$S_{n,1} := \mathbb{E}\left[\left(\sum_{\xi\in\mathbb{Z}^d, \|h_n\xi\|\le 1} \|\xi\|^{-2}\left|\mathcal{F}[\widetilde{p}_n - p_{h_n}](\xi)\right|^2\right)^{\frac{\rho}{2}}\right],$$

$$S_{n,2} := \mathbb{E}\left[\left(\sum_{\xi\in\mathbb{Z}^d, \|h_n\xi\| > 1} \|\xi\|^{-2}\left|\mathcal{F}[\widetilde{p}_n - p_{h_n}](\xi)\right|^2\right)^{\frac{\rho}{2}}\right].$$

We begin by bounding $S_{n,1}$. Recall that

$$\mathcal{F}[\widetilde{p}_n - p_{h_n}](\xi) = \mathcal{F}[K](h_n\xi)\frac{1}{n}\sum_{j=1}^n \left(e^{-2\pi i\langle X_j,\xi\rangle} - \mathcal{F}[p](\xi)\right), \quad \xi\in\mathbb{Z}^d,$$

where $i^2 = -1$. In fact, since $\widetilde{p}_n$ and p_{h_n} integrate to the same constant, we have $\mathcal{F}[\widetilde{p}_n - p_{h_n}](0) = 0$. Furthermore, let $\rho'\in\mathbb{R}$ satisfy $\frac{1}{\rho}+\frac{1}{\rho'} = \frac{1}{2}$. Then, for any $\eta\in\mathbb{R}$, we have by Hölder's inequality,

$$\begin{aligned}
S_{n,1} &= \mathbb{E}\left[\left(\sum_{\xi\in\mathbb{Z}^d, \|h_n\xi\|\le 1} \|\xi\|^{-2\eta}\|\xi\|^{2(\eta-1)}\left|\mathcal{F}[\widetilde{p}_n - p_{h_n}](\xi)\right|^2\right)^{\frac{\rho}{2}}\right] \\
&\le \left(\sum_{\substack{\xi\in\mathbb{Z}^d,\xi\neq 0\\ \|h_n\xi\|\le 1}} \|\xi\|^{-\rho'\eta}\right)^{\frac{\rho}{\rho'}} \mathbb{E}\left[\sum_{\substack{\xi\in\mathbb{Z}^d,\xi\neq 0\\ \|h_n\xi\|\le 1}} \|\xi\|^{\rho(\eta-1)}\left|\mathcal{F}[\widetilde{p}_n - p_{h_n}](\xi)\right|^{\rho}\right] \\
&= \left(\sum_{\substack{\xi\in\mathbb{Z}^d,\xi\neq 0\\ \|h_n\xi\|\le 1}} \|\xi\|^{-\rho'\eta}\right)^{\frac{\rho}{\rho'}} \sum_{\substack{\xi\in\mathbb{Z}^d,\xi\neq 0\\ \|h_n\xi\|\le 1}} \|\xi\|^{\rho(\eta-1)}\left|\mathcal{F}[K](h_n\xi)\right|^{\rho}\mathbb{E}\left|\frac{1}{n}\sum_{j=1}^n Z_j(\xi)\right|^{\rho},
\end{aligned}$$

where $Z_j(\xi) = e^{-2\pi i\langle X_j,\xi\rangle} - \mathcal{F}[p](\xi)$, for all $j=1,\dots,n$ and $\xi\in\mathbb{Z}^d$. Since $\rho\ge 2$, it can be deduced from Rosenthal's inequalities (Rosenthal, 1970, 1972) that,

$$\mathbb{E}\left|\frac{1}{n}\sum_{j=1}^n Z_j(\xi)\right|^{\rho} \lesssim n^{-\frac{\rho}{2}}\left(\mathbb{E}|Z_1(\xi)|^2\right)^{\rho} + n^{1-\rho}\mathbb{E}|Z_1(\xi)|^{\rho}.$$

Notice that $|Z_1(\xi)|\le 2$ for any $\xi\in\mathbb{Z}^d$, and $\rho/2\le\rho-1$, thus we deduce from the previous two displays that,

$$S_{n,1} \lesssim n^{-\frac{\rho}{2}}\left(\sum_{\substack{\xi\in\mathbb{Z}^d,\xi\neq 0\\ \|h_n\xi\|\le 1}} \|\xi\|^{-\rho'\eta}\right)^{\frac{\rho}{\rho'}} \sum_{\substack{\xi\in\mathbb{Z}^d,\xi\neq 0\\ \|h_n\xi\|\le 1}} \|\xi\|^{\rho(\eta-1)}\left|\mathcal{F}[K](h_n\xi)\right|^{\rho} \tag{5.60}$$

$$\lesssim n^{-\frac{\rho}{2}} \left(\sum_{\substack{\xi \in \mathbb{Z}^d, \xi \neq 0 \\ \|h_n \xi\| \leq 1}} \|\xi\|^{-\rho' \eta} \right)^{\frac{\rho}{\rho'}} \sum_{\substack{\xi \in \mathbb{Z}^d, \xi \neq 0 \\ \|h_n \xi\| \leq 1}} \|\xi\|^{\rho(\eta - 1)}, \tag{5.61}$$

where the final inequality follows from the fact that the Fourier transform of K is bounded over the unit ball, since $K \in \mathcal{C}_c^\infty(\mathbb{R}^d)$. When $d \geq 3$, due to the condition $\frac{1}{\rho} + \frac{1}{\rho'} = \frac{1}{2}$, we may choose η satisfying

$$d\left(\frac{1}{d} - \frac{1}{\rho}\right) < \eta < \frac{d}{\rho'}. \tag{5.62}$$

In particular, we then have $-d < \rho(\eta - 1)$ and $-d < -\rho'\eta$, so that

$$S_{n,1} \lesssim n^{-\frac{\rho}{2}} \left(h_n^{\rho'\eta - d}\right)^{\frac{\rho}{\rho'}} h_n^{-\rho(\eta-1)-d} = n^{-\frac{\rho}{2}} h_n^{\rho - d(\frac{\rho}{\rho'}+1)} = n^{-\frac{\rho}{2}} h_n^{\rho(1-\frac{d}{2})}.$$

If $d = 2$, we choose η such that the strict inequalities in equation (5.62) both hold with equality. In this case, we have $\rho'\eta = d$ and $\rho(\eta - 1) = -d$, thus

$$S_{n,1} \lesssim n^{-\frac{\rho}{2}} \left(\sum_{\xi \in \mathbb{Z}^d, \|h_n \xi\| \leq 1} \|\xi\|^{-d} \right)^{\frac{\rho}{\rho'}+1} \lesssim n^{-\frac{\rho}{2}} \log(h_n^{-1})^{\frac{\rho}{\rho'}+1} = \left(\log(h_n^{-1})/n\right)^{\frac{\rho}{2}}.$$

Finally, if $d = 1$, choose η such that

$$1 - \frac{1}{\rho} > \eta > \frac{1}{\rho'}. \tag{5.63}$$

In this case, both sequences in equation (5.61) are summable, and we obtain $S_{n,1} \lesssim n^{-\rho/2}$. In summary, we deduce

$$S_{n,1} \lesssim \beta_n := n^{-\frac{\rho}{2}} \begin{cases} h_n^{\rho(1-\frac{d}{2})}, & d \geq 3 \\ \left(\log(h_n^{-1})\right)^{\rho/2}, & d = 2 \\ 1, & d = 1. \end{cases} \tag{5.64}$$

We next bound $S_{n,2}$. Let $\eta < d/\rho'$. Apply a similar reduction as in equation (5.60), to obtain

$$S_{n,2} \lesssim n^{-\frac{\rho}{2}} \left(\sum_{\xi \in \mathbb{Z}^d, \|h_n \xi\| > 1} \|\xi\|^{-\rho'\eta} \right)^{\frac{\rho}{\rho'}} \left(\sum_{\xi \in \mathbb{Z}^d, \|h_n \xi\| > 1} \|\xi\|^{\rho(\eta-1)} \left|\mathcal{F}[K](h_n\xi)\right|^\rho \right).$$

Since $K \in \mathcal{C}_c^\infty(\mathbb{R}^d)$, notice that K and $\mathcal{F}[K]$ are Schwartz functions. In particular, $\mathcal{F}[K](\xi) \lesssim \|\xi\|^{-\ell}$ for any $\ell > 0$. Choose $\ell > 0$ such that $\rho(\eta - 1 - \ell) < -d$. We then have,

$$S_{n,2} \lesssim n^{-\frac{\rho}{2}} \left(\sum_{\xi \in \mathbb{Z}^d, \|h_n \xi\| > 1} \|\xi\|^{-\rho'\eta} \right)^{\frac{\rho}{\rho'}} \left(h_n^{-\rho\ell} \sum_{\xi \in \mathbb{Z}^d, \|h_n \xi\| > 1} \|\xi\|^{\rho(\eta-1-\ell)} \right)$$

$$\lesssim n^{-\frac{\rho}{2}} \left(h_n^{\rho'\eta-d}\right)^{\frac{\rho}{\rho'}} h_n^{-\rho\ell} h_n^{-\rho(\eta-1-\ell)-d} \lesssim n^{-\frac{\rho}{2}} h_n^{\rho(1-\frac{d}{2})} \lesssim \beta_n.$$

Combine this bound with those of equations (5.59) and (5.64)

$$\mathbb{E}\mathcal{W}_2^\rho(\widehat{P}_n, P) \lesssim h_n^{\rho(s+1)} + \beta_n + 1/n^2 \lesssim \begin{cases} n^{-\frac{\rho(s+1)}{2s+d}}, & d \geq 3 \\ (\log n/n)^{\rho/2}, & d = 2 \\ (1/n)^{\rho/2}, & d = 1. \end{cases}$$

The claim follows. □

We are now in a position to prove Theorem 20.

5.H.1 Proof of Theorem 20

In view of Propositions 42–43 and Lemma 46, the proof of the claim is analogous to that of Theorem 22, thus we only provide brief justifications.

Regarding part (i), apply Lemmas 102 and Proposition 42 to deduce that there exists $\epsilon \in (0, 1 \wedge \frac{\alpha-1}{2})$ and an event of probability at least $1 - 1/n^2$ over which $\widehat{p}_n, \widehat{q}_m$ coincide with $\widetilde{p}_n, \widetilde{q}_m$ respectively, are bounded from below by $(2\gamma)^{-1}$, and are of class $\mathcal{C}^\epsilon(\mathbb{T}^d)$, with Hölder norm uniformly bounded in n. By Theorem 4, it follows that, over this same high-probability event, any mean-zero Brenier potential in the optimal transport problem from P to $\widehat{Q}_m$, or from $\widehat{P}_n$ to $\widehat{Q}_m$, is of class $\mathcal{C}^{2+\epsilon}(\mathbb{T}^d)$, again with a uniformly bounded Hölder norm. Arguing as in Step 1 of the proof of Theorem 22(i), we deduce that these potentials achieve the conclusion of equation (5.54) therein. The same argument as in Steps 2–3 of that proof, coupled with Lemma 46 stating the convergence rate of the kernel density estimator in Wasserstein distance, can then be used to deduce that the optimal transport map $\widehat{T}_{nm}$ from $\widehat{P}_n$ to $\widehat{Q}_m$ satisfies

$$\mathbb{E}\big\|\widehat{T}_{nm} - \widehat{T}_0\big\|_{L^2(P)}^2 \lesssim \mathbb{E}W_2^2(\widehat{P}_n, P) + \mathbb{E}W_2^2(\widehat{Q}_m, Q) + \frac{1}{(n \wedge m)^2} \lesssim R_{K, n\wedge m}(\alpha).$$

In applying Lemma 46, we note that our stated assumption $\mathbf{K(2\alpha)}$ implies $\mathbf{K(\alpha + 1)}$ for a constant $\kappa' > 0$ depending only on α and κ. This proves part (i). To prove part (ii), we use the following observation.

Lemma 47. *Under the assumptions of Theorem 20, we have*

$$\mathbb{E}\left[\int \phi_0(\widehat{p}_n - p)\right] = O(h_n^{2\alpha}), \quad \mathrm{Var}\left[\int \phi_0(\widehat{p}_n - p)\right] = \frac{\mathrm{Var}_P[\phi_0(X)]}{n} + O\left(\frac{h_n^{2\alpha}}{n}\right), \tag{5.65}$$

where the implicit constants depend only on M, γ, α.

Using Lemmas 46–47, the same argument as in the proof of Theorem 22(ii) leads to the claim of part (ii). □

It thus remains to prove Lemma 47.

5.H.2 Proof of Lemma 47

Using Proposition 42, the densities $\widetilde{p}_n$ and $\widehat{p}_n$ coincide with high probability, thus arguing similarly as in the proof of Lemma 36, it will suffice to prove that

$$\int \phi_0(p - p_{h_n}) = O(h_n^{2\alpha}), \quad \mathrm{Var}\left[\int \phi_0(\widetilde{p}_n - p_{h_n})\right] = \frac{\mathrm{Var}_P[\phi_0(X)]}{n} + O\left(\frac{h_n^{2\alpha}}{n}\right). \tag{5.66}$$

Under the condition $\alpha \notin \mathbb{N}$, $\alpha > 1$, we deduce from Theorem 4 that there exists $\lambda > 0$ depending only on M, γ, α such that

$$\phi_0, \psi_0 \in \mathcal{C}^{\alpha+1}(\mathbb{T}^d; \lambda). \tag{5.67}$$

Now, by Parseval's Theorem,

$$\begin{aligned}\left|\int_{\mathbb{T}^d} \phi_0(p - p_{h_n})\right| &= \left|\sum_{\xi \in \mathbb{Z}^d} \mathcal{F}[\phi_0](\xi)\mathcal{F}[p - p_{h_n}](\xi)\right| \\ &\leq \big\| \|\cdot\|^{\alpha+1}\mathcal{F}[\phi_0](\cdot)\big\|_{\ell^2(\mathbb{Z}^d)} \big\| \|\cdot\|^{-(\alpha+1)}\mathcal{F}[p - p_{h_n}](\cdot)\big\|_{\ell^2(\mathbb{Z}^d)} \\ &= \|\phi_0\|_{\dot{H}^{\alpha+1}(\mathbb{T}^d)} \|p - p_{h_n}\|_{\dot{H}^{-(\alpha+1)}(\mathbb{T}^d)} \lesssim \|p - p_{h_n}\|_{\dot{H}^{-(\alpha+1)}(\mathbb{T}^d)},\end{aligned}$$

where we used equation (5.67) and the fact that $\|\phi_0\|_{\dot{H}^{\alpha+1}(\mathbb{T}^d)} \leq \|\phi_0\|_{H^{\alpha+1}(\mathbb{T}^d)} \lesssim \|\phi_0\|_{\mathcal{C}^{\alpha+1}(\mathbb{T}^d)}$. Apply Proposition 43, under the assumption $\mathbf{K(2\alpha)}$ and the smoothness assumption on p, to deduce

$$\left|\int_{\mathbb{T}^d} \phi_0 d(p - p_{h_n})\right| \lesssim h_n^{2\alpha} \lesssim R_{K,n\wedge m}(\alpha).$$

To prove the variance bound, notice that

$$\mathrm{Var}\left[\int \phi_0(\widetilde{p}_n - p_{h_n})\right] = \mathrm{Var}\left[\int (\phi_0 \star K_{h_n}) d(P_n - P)\right] = \frac{1}{n}\mathrm{Var}_P[\phi_{h_n}(X)],$$

where $\phi_{h_n} = \phi_0 \star K_{h_n}$. Thus, reasoning as in the proof of Lemma 36, we have

$$\begin{aligned}&\left|\mathrm{Var}\left[\int \phi_0(\widetilde{p}_n - p_{h_n})\right] - \frac{1}{n}\mathrm{Var}_P[\phi_0(X)]\right| \\ &\quad\leq \frac{1}{n}\Big|\mathrm{Var}_P[\phi_{h_n}(X)] - \mathrm{Var}_P[\phi_0(X)]\Big| \\ &\quad\leq \frac{1}{n}\Big|\mathbb{E}[\phi_{h_n}^2(X) - \phi_0^2(X)]\Big| + \frac{1}{n}\Big|\mathbb{E}[\phi_{h_n}(X) - \phi_0(X)]\Big| = \frac{1}{n}\big[(I) + (II)\big].\end{aligned}$$

We shall again bound term (I), and a similar proof can be used for term (II). Notice that

$$(I) = \left|\int (\phi_{h_n} - \phi_0)(\phi_{h_n} + \phi_0)p\right| \leq \|\phi_{h_n} - \phi_0\|_{\dot{H}^{-(\alpha-1)}(\mathbb{T}^d)} \|(\phi_{h_n} + \phi_0)p\|_{\dot{H}^{\alpha-1}(\mathbb{T}^d)}.$$

It is a straightforward observation that $\|\phi_{h_n}\|_{\mathcal{C}^{\alpha+1}(\mathbb{T}^d)} \leq \|\phi_0\|_{\mathcal{C}^{\alpha+1}(\mathbb{T}^d)}$ for all $n \geq 1$, thus the function $(\phi_{h_n} + \phi_0)p$ has uniformly bounded $\mathcal{C}^{\alpha-1}(\mathbb{T}^d)$ norm, by Lemma 95. Since $\phi_0 \in \mathcal{C}^{\alpha+1}(\mathbb{T}^d; \lambda)$, we deduce that

$$(I) \lesssim \|\phi_{h_n} - \phi_0\|_{\dot{H}^{-(\alpha-1)}(\mathbb{T}^d)} \lesssim h_n^{2\alpha},$$

by Proposition 43. The claim follows from here. □

5.H.3 Further Results

In this section, we state for completeness several additional results on estimating optimal transport maps and Wasserstein distances over $\mathbb{T}^d$, which mirror our results over domains of $\mathbb{R}^d$ across Sections 5.2–5.3. Throughout what follows, let $P, Q \in \mathcal{P}_{\mathrm{ac}}(\mathbb{T}^d)$ admit densities p, q, and let $X_1, \dots, X_n \sim P$ and $Y_1, \dots, Y_m \sim Q$ be i.i.d. samples which are independent of each other. Let $P_n = (1/n)\sum_{i=1}^n \delta_{X_i}$ and $Q_m = (1/m)\sum_{j=1}^m \delta_{Y_j}$. As in the previous subsections, we omit the superscript "ker" on the estimators $\widehat{P}_n^{(\mathrm{ker})}$ and $\widehat{Q}_m^{(\mathrm{ker})}$. Let T_m be the optimal transport map from P to Q_m, and let

$$\Delta_{nm} = \sum_{i=1}^n \sum_{j=1}^m \widehat{\pi}_{ij} \|T_0(X_i) - Y_j\|^2,$$

where

$$\widehat{\pi} \in \operatorname*{argmin}_{\pi \in \mathcal{Q}_{nm}} \sum_{i=1}^n \sum_{j=1}^m \pi_{ij} d_{\mathbb{T}^d}^2(X_i, Y_j).$$

Furthermore, let $\overline{T}_m$ be the optimal transport map from P to $\widehat{Q}_m$. We begin by stating a one-sample analogue of Theorem 20.

Proposition 30 (One-Sample Kernel Estimators). Let $P, Q \in \mathcal{P}_{\mathrm{ac}}(\mathbb{T}^d)$ and assume $p, q \in \mathcal{C}^{\alpha-1}(\mathbb{T}^d; M, \gamma)$, for some $\alpha > 1$, $\alpha \notin \mathbb{N}$, and $M, \gamma > 0$. Let $h_m \asymp m^{-1/(d+2(\alpha-1))}$. Then, there exists a constant $C > 0$ depending only on M, γ, α such that the following assertions hold.

(i) (Optimal Transport Maps) We have,

$$\mathbb{E}\big\|\overline{T}_m - T_0\big\|_{L^2(P)}^2 \leq C R_{K,m}(\alpha).$$

(ii) (Wasserstein Distances) We have,

$$\begin{aligned}\big|\mathbb{E}W_2^2(P, \widehat{Q}_m) - W_2^2(P,Q)\big| &\leq C R_{K,m}(\alpha),\\ \mathbb{E}\big|W_2^2(P, \widehat{Q}_m) - W_2^2(P,Q)\big|^2 &\leq \left[C R_{K,m}(\alpha) + \sqrt{\frac{\mathrm{Var}_Q[\psi_0(Y)]}{m}}\right]^2.\end{aligned}$$

Next, we state convergence rates for empirical estimators. In what follows, we use the abbreviation

$$\widetilde{\kappa}_n = \begin{cases} 1/n, & d = 1, \\ \log n/n, & d = 2, \\ n^{-2/d}, & d \geq 3. \end{cases}$$

Proposition 31 (One-Sample Empirical Estimators over $\mathbb{T}^d$)**.** Let $P, Q \in \mathcal{P}_{\mathrm{ac}}(\mathbb{T}^d)$ admit densities p, q satisfying

$$\gamma^{-1} \leq p, q \leq \gamma, \quad \text{over } \mathbb{T}^d,$$

for some $\gamma > 0$. Assume further that $\phi_0 \in \mathcal{C}^2(\mathbb{T}^d)$. Then,

$$\mathbb{E}\|\widehat{T}_m - T_0\|_{L^2(P)}^2 \asymp \mathbb{E}\big[W_2^2(P, Q_m) - W_2^2(P, Q)\big] \asymp \mathbb{E}W_2^2(Q_m, Q) \lesssim \widetilde{\kappa}_m,$$

and,

$$\mathbb{E}[\Delta_{nm}] \asymp \mathbb{E}\big[W_2^2(P_n, Q_m) - W_2^2(P, Q)\big] \lesssim \widetilde{\kappa}_{n \wedge m}.$$

From Proposition 31, one may also deduce rates of convergence for the nearest-neighbor estimator discussion in Section 5.3.2, over $\mathbb{T}^d$. We omit the details for the sake of brevity.

5.I Plugin Estimation over Smooth Domains

The goal of this appendix is to prove Theorem 21. Let us summarize our proof strategy.

(i) In Section 5.I.1, we define a scale of spectrally-defined Sobolev spaces $\mathcal{H}^{s,r}(\Omega)$, which are well-suited to the analysis of the density estimator $\widehat{q}_n^{(\mathrm{lap})}$. We show that these spaces coincide with a scale of subspaces $H_N^{s,r}(\Omega)$ of the usual Sobolev spaces $H^{s,r}(\Omega)$.

(ii) In Section 5.I.2, we bound the risk of the density estimator $\widehat{q}_n^{(\mathrm{lap})}$ under the norm of the space $\mathcal{H}^{s,r}(\Omega)$, for a wide range of parameters s, r.

(iii) In Section 5.I.3, we use parts (i)–(ii) to derive a convergence rate of $\widehat{q}_n^{(\mathrm{lap})}$ under a suitable Hölder norm. This will allow us to deduce that with probability tending to one, $\widehat{q}_n^{(\mathrm{lap})}$ satisfies the regularity conditions needed for Caffarelli's regularity theory (condition **(C2)**).

(iv) In Section 5.I.4, we combine parts (i)–(iii) to obtain a convergence rate of $\widehat{q}_n^{(\mathrm{lap})}$ under the Wasserstein distance, using its equivalence to a negative Sobolev norm (cf. equation (5.9)).

(v) In Section 5.I.5, we combine these ingredients to deduce the claim, by the same strategy as in the proof of Theorems 22 and 20.

Throughout our development, an important role will be played by the Neumann boundary condition, together with assumption **(C1)**. On the one hand, we will see that these conditions are sufficient for the space $\mathcal{H}^{s,r}(\Omega)$ to admit a Littlewood-Paley characterization; cf. Lemmas 48 and 51. On the other hand, these conditions ensure that the eigenvalues and spectral function of the Neumann Laplacian grow at a sufficiently slow rate to obtain our stated convergence rates for density estimation; cf. Lemmas 53–54.

5.I.1 Spectrally-Defined Sobolev Spaces

To facilitate our analysis of the estimators $\widetilde{p}_n, \widetilde{q}_m$, we will begin by showing that the Bessel potential Sobolev spaces $H^{s,r}(\Omega)$, defined in Appendix A, can be characterized via the spectrum of the Neumann Laplacian. More specifically, we will work with the following subspaces of $H^{s,r}(\Omega)$, which have suitably vanishing Neumann trace. Throughout what follows, we always denote by ν an outward-pointing unit normal vector to $\partial\Omega$, and by $\partial u/\partial\nu$ the weak normal derivative operator, whose trace on $\partial\Omega$ is well-defined and takes values in $L^r(\partial\Omega)$ whenever $u \in H^{t,r}(\Omega)$ with $t > 1 + 1/r$ (Taira, 2016, Theorem 4.6). We will assume $s \geq 0$ and $r \geq 2$ throughout this section. Furthermore, we adopt the nonstandard notation $H_0^{s,r}(\Omega) = H^{s,r}(\Omega) \cap L_0^r(\Omega)$.

Definition 2 (Triebel (1995), Section 4.3.3)**.** Let Ω be a domain satisfying condition **(C1)**. For all $s \geq 0$ and $r \geq 2$, let $H_N^{s,r}(\Omega)$ be defined as follows.

(i) If $s - 1/r < 1$, set $H_N^{s,r}(\Omega) = H_0^{s,r}(\Omega)$.

(ii) If $2k + 1 < s - 1/r < 2(k+1) + 1$ for some $k \geq 0$, set

$$H_N^{s,r}(\Omega) = \left\{ u \in H_0^{s,r}(\Omega) : \frac{\partial \Delta^j u}{\partial \nu} = 0 \text{ on } \partial\Omega,\ 0 \leq j \leq k \right\}.$$

(iii) If $2k + 1 = s - 1/r$ for some $k \geq 0$, extend ν continuously to Ω, and set

$$H_N^{s,r}(\Omega) = \left\{ u \in H_0^{s,r}(\Omega) : \frac{\partial \Delta^j u}{\partial \nu} = 0 \text{ on } \partial\Omega,\ 0 \leq j < k,\ \frac{\partial \Delta^k u}{\partial \nu} \in H^{\frac{1}{r},r}(\mathbb{R}^d) \right\},$$

where $\partial\Delta^k u/\partial\nu$ is extended by zero outside of Ω.

Note that, in part (iii), the normal vector ν may be extended smoothly away from the boundary since we assumed $\partial\Omega$ is $\mathcal{C}^\infty$, thus ν is itself smooth. The particular choice of extension does not alter the definition of the space. As we shall see, the relevance of the space $H_N^{s,r}(\Omega)$ lies in the fact that the fractional Neumann Laplacian $(-\Delta)^{s/2}$, defined next, is an isomorphism of $H_N^{s,r}(\Omega)$ onto $L_0^r(\Omega)$.

Recall that $0 < \lambda_1 \leq \lambda_2 \leq \dots$ is the sequence of eigenvalues corresponding to the eigenbasis $\{\eta_\ell\}_{\ell=1}^\infty$. Define the spectral fractional Laplacian for all $u \in L_0^2(\Omega)$ by

$$(-\Delta)^{s/2} u = \sum_{\ell=1}^\infty \lambda_\ell^{s/2} \mathcal{L}_\ell[u] \eta_\ell, \quad \text{where } \mathcal{L}_\ell[u] := \langle u, \eta_\ell \rangle_{L^2(\Omega)},\ \ \ell = 1, 2, \dots$$

Furthermore, for $s \geq 0$ and $r \geq 2$, let $\mathcal{H}^{s,r}(\Omega)$ denote the Banach space of functions $u \in L_0^r(\Omega)$ such that the norm

$$\|u\|_{\mathcal{H}^{s,r}(\Omega)} := \left\|(-\Delta)^{s/2} u\right\|_{L^r(\Omega)} < \infty.$$

is finite. In the special case $r = 2$, $\mathcal{H}^{s,r}(\Omega)$ becomes a Hilbert space, as noted by Dunlop et al. (2020). In this case we omit the superscript "r" and simply write $\mathcal{H}^s(\Omega) := \mathcal{H}^{s,2}(\Omega)$. The corresponding inner product on this space is given by

$$\langle u, v\rangle_{\mathcal{H}^s(\Omega)} = \sum_{\ell=1}^{\infty} \lambda_\ell^s \mathcal{L}_\ell[u]\mathcal{L}_\ell[v], \quad u, v \in \mathcal{H}^s(\Omega).$$

It is easy to see by Parseval's identity that $\|\cdot\|_{\mathcal{H}^s(\Omega)} = \sqrt{\langle \cdot, \cdot\rangle_{\mathcal{H}^s(\Omega)}}$.

Our first aim is to show that the spaces $\mathcal{H}^{s,r}(\Omega)$ and $H_N^{s,r}(\Omega)$ coincide, with equivalent norms. To simplify our proof, we will focus only on the ranges of s and r which we will need in our development.

Proposition 32. Let Ω satisfy condition **(C1)**. Assume that one of the following conditions holds:

$$\begin{cases} s \in [0,2] \\ r \geq 2 \end{cases} \quad \text{or} \quad \begin{cases} s \geq 0 \\ r = 2. \end{cases} \tag{5.68}$$

Then, with equivalent norms,

$$H_N^s(\Omega) = \mathcal{H}^s(\Omega). \tag{5.69}$$

Proposition 32 was stated without proof by Seeley (1972, Section 5). For completeness, we provide a self-contained proof below. Let us also note that, in the Hilbertian case $r = 2$, Proposition 32 was established for integer exponents s by Dunlop et al. (2020, Lemma 7.1), and for $s \in [0,2]$ by Kim, Balakrishnan, and Wasserman (2020). For the case $r > 2$, a result similar to Proposition 32 was proven by Cao and Grigor'yan (2020, Theorem 3.1), however they considered the setting where Ω is the entire Euclidean space $\mathbb{R}^d$, thus they employed a different definition of the fractional Laplacian.

Before turning to the proof, let us begin by stating a generalization of Mikhlin's multiplier theorem for the Neumann Laplacian, which we will use repeatedly in the following subsections. This result follows from Theorem 1.3 of Xu (2011), or Theorem 7.9 of Kerkyacharian and Petrushev (2015).

Lemma 48 (Mikhlin's Multiplier Theorem for the Neumann Laplacian)**.** *Let $m \in \mathcal{C}^\infty(\mathbb{R}_+)$ satisfy Mikhlin's multiplier condition:*

$$|D^\alpha m(x)| \leq c|x|^{-|\alpha|}, \quad \textit{for all } \alpha = 1, \dots, d+1, \; x \in \mathbb{R}_+. \tag{5.70}$$

Then, for any $1 < r < \infty$, there exists $C > 0$ depending on Ω, c, r such that for any $f \in L_N^r(\Omega)$,

$$\left\| \sum_{\ell=1}^{\infty} m(\sqrt{\lambda_\ell})\mathcal{L}_\ell[f]\eta_\ell \right\|_{L^r(\Omega)} \leq C\|f\|_{L^r(\Omega)}. \tag{5.71}$$

In this case, we say that m is an $L_N^r(\Omega)$ multiplier.

We now turn to the proof of Proposition 32.

5.I.1.1 Proof of Proposition 32

Let us begin with the case where $s \in [0,2]$ and $r \geq 2$. The result is trivial when $s = 0$, in which case we have

$$\mathcal{H}^{0,r}(\Omega) = H_N^{0,r}(\Omega) = L_0^r(\Omega).$$

Next, we prove the claim when $s = 2$, in which case the space $H_N^{s,r}(\Omega)$ takes the form

$$H_N^{2,r}(\Omega) = \left\{ u \in H^{2,r}(\Omega) : \frac{\partial u}{\partial \nu} = 0 \text{ on } \partial\Omega \right\}.$$

By Triebel (1995, Theorem 4.2.4, p. 316), we have that $H^{2,s}(\Omega) = W^{2,s}(\Omega)$ with equivalent norms, where $W^{k,r}(\Omega)$ is the standard $L^r(\Omega)$ Sobolev norm with integer smoothness parameter $k \in \mathbb{N}$. Thus, for all $u \in H_N^{2,r}(\Omega)$, we have

$$\|u\|_{\mathcal{H}^{2,r}(\Omega)} = \|\Delta u\|_{L^r(\Omega)} \lesssim \|u\|_{W^{2,r}(\Omega)} \lesssim \|u\|_{H^{2,r}(\Omega)}.$$

It is a standard fact that $-\Delta$ is a bijection of $H_N^{2,r}(\Omega)$ onto $L_0^r(\Omega)$ (e.g. Franke and Runst (1995)). Furthermore, the above display shows that this mapping is continuous, thus, by the Banach isomorphism theorem, the operator $(-\Delta)^{-1} : L_0^r(\Omega) \to H_N^{2,r}(\Omega)$ is bounded. Equivalently, for all $w \in \mathcal{H}^{2,r}(\Omega)$, we obtain $\|w\|_{H^{2,r}(\Omega)} \lesssim \|w\|_{\mathcal{H}^{2,r}(\Omega)}$. We deduce that $\mathcal{H}^{2,r}(\Omega) = H_N^{2,r}(\Omega)$, with equivalent norms.

It thus remains to prove the claim for all $s \in (0,2)$, which we shall do using an interpolation argument. Given two complex Banach spaces A and B, let $(A,B)_{[\theta]}$ denote the complex interpolation space between A and B, for any $\theta \in [0,1]$ (Bergh and Löfström, 1976). It is well-known that the complex interpolation of any two Bessel potential spaces $H^{s_0,r}(\Omega)$ and $H^{s_1,r}(\Omega)$ is itself a Bessel potential space (see for instance Triebel (1995), Theorem 4.3.1/1). The following is an analogue of this result for spaces with zero Neumann trace.

Lemma 49 (Seeley (1972)). *Let $1 < r < \infty$. Then, for all $s \geq 0$ and $\theta \in (0,1)$,*

$$H_N^{\theta s,r}(\Omega) = \big(L_0^r(\Omega), H_N^{s,r}(\Omega)\big)_{[\theta]}.$$

By combining Lemma 49 with what we have shown above, it holds for any $s \in [0,2]$,

$$H_N^{s,r}(\Omega) = \big(L_0^r(\Omega), H_N^{2,r}(\Omega)\big)_{[s/2]} = \big(L_0^r(\Omega), \mathcal{H}^{2,r}(\Omega)\big)_{[s/2]},$$

To complete the proof of the claim, it thus suffices to prove that

$$\mathcal{H}^{s,r}(\Omega) = \big(L_0^r(\Omega), \mathcal{H}^{2,r}(\Omega)\big)_{[s/2]} \tag{5.72}$$

for any $s \in [0,2]$. We will do so by following similar lines as the proof of Theorem 6.4.5 of Bergh and Löfström (1976). Specifically, the following can be inferred from their Theorem 6.4.2.

Lemma 50. *Suppose there exists a collection of complex Banach spaces $(B_s)_{s\in[0,2]}$ with $B_s \subseteq B_{s'} \subseteq L_N^r(\Omega)$ for all $0 \leq s' \leq s \leq 2$, fulfilling the following properties for all $s \in [0,2]$:*

(i) $B_s = (B_0, B_2)_{[s/2]}$.

(ii) There exists a continuous linear map $\mathscr{I} : \mathcal{H}^{s,r}(\Omega) \to B_s$.

(iii) There exists a continuous linear map $\mathscr{P} : B_s \to \mathcal{H}^{s,r}(\Omega)$ *such that* $\mathscr{P} \circ \mathscr{I} = \mathrm{Id}_{\mathcal{H}^{s,r}(\Omega)}$.

Then, equation (5.72) *holds for all* $s \in [0, 2]$.

The claim will therefore follow if we can exhibit a collection of Banach spaces $(B_s)_{s\in[0,2]}$ satisfying the properties of Lemma 50. To do so, it will be convenient to show that $\mathcal{H}^{s,r}(\Omega)$ lies in the Triebel-Lizorkin family of spaces. Indeed, it is well-known that the standard Sobolev space $H^{s,r}(\Omega)$ is equal to the Triebel-Lizorkin space $F^s_{r,2}(\Omega)$ (Triebel, 1995), and an analogue of this fact for the space $\mathcal{H}^{s,r}(\Omega)$ has been derived by Kerkyacharian and Petrushev (2015). We state a variant of their result below, beginning with some notation. Let $\xi_0, \xi \in \mathcal{C}^\infty(\mathbb{R}_+)$ be an admissible pair of Littlewood-Paley functions, so that $\mathrm{supp}(\xi_0) \subseteq [0, 2]$, $\mathrm{supp}(\xi) \subseteq [1/2, 2]$, and $\sum_{j\geq 0} \xi_j(\lambda) = 1$ for all $\lambda \in \mathbb{R}$, where we write $\xi_j = \xi(2^{-j}(\cdot))$ (cf. Lemma 6.1.7 of Bergh and Löfström (1976) for a construction of such functions). We then have the following statement.

Lemma 51 (Kerkyacharian and Petrushev (2015))**.** *Let* $1 < r < \infty$. *Then, for all* $u \in L^r_0(\Omega)$,

$$\|u\|_{\mathcal{H}^{s,r}(\Omega)} \asymp \|u\|_{\mathscr{F}^s_{r,2}(\Omega)} := \left\| \left(\sum_{j=0}^\infty \left| 2^{js} \sum_{\ell=1}^\infty \xi_j(\lambda_\ell^{1/2}) \mathcal{L}_\ell[u] \eta_\ell(\cdot) \right|^2 \right)^{\frac{1}{2}} \right\|_{L^r(\Omega)}$$

Proof of Lemma 51. It follows from Theorem 7.8 of Kerkyacharian and Petrushev (2015) that

$$\|u\|_{\mathscr{F}^s_{r,2}(\Omega)} \asymp \left\| (\mathrm{Id} - \Delta)^{s/2} u \right\|_{L^r(\Omega)} = \left\| \sum_{\ell=1}^\infty (1 + \lambda_\ell)^{s/2} \mathcal{L}_\ell[u] \eta_\ell \right\|_{L^r(\Omega)}$$

where we used the fact the Neumann Laplacian over a convex domain Ω with smooth boundary satisfies the conditions of the operator L in the introduction of Kerkyacharian and Petrushev (2015). Indeed, the Gaussian upper bounds on the heat kernel generated by the Neumann Laplacian are given for instance in Theorem 3.3.5 of Davies (1989), while the Hölder continuity of the heat kernel can be deduced, as in Proposition 3.1 of Sturm (1996), from the parabolic Harnack inequality (see for instance Theorem 5.3.5 of Davies (1989)). See also Saloff-Coste (2010, Theorem 3.1), and remarks thereafter. To prove our claim, it thus suffices to show that

$$\|(-\Delta)^{s/2} u\|_{L^r(\Omega)} \asymp \|(\mathrm{Id} - \Delta)^{s/2} u\|_{L^r(\Omega)}. \tag{5.73}$$

Notice first that the maps $m(\lambda) = (1 + \lambda^{s/2})/(1 + \lambda)^{s/2}$ and $1/m(\lambda)$, $\lambda \in \mathbb{R}_+$ satisfy the conditions of Lemma 48, thus

$$\|(\mathrm{Id} - \Delta)^{s/2} u\|_{L^r(\Omega)} \asymp \left\| \sum_{\ell=1}^\infty (1 + \lambda_\ell^{s/2}) \mathcal{L}_\ell[u] \eta_\ell \right\|_{L^r(\Omega)} \asymp \|u\|_{L^r(\Omega)} + \|(-\Delta)^{s/2} u\|_{L^r(\Omega)}.$$

It thus suffices to show that $\|u\|_{L^r(\Omega)} \lesssim \|(-\Delta)^{s/2}u\|_{L^r(\Omega)}$. This follows from the fact that any map $m \in \mathcal{C}^\infty(\mathbb{R}_+)$, of the form $m(\lambda) = \lambda^{-s/2}$ for $\lambda > \lambda_1/2$, is an $L^r_N(\Omega)$ multiplier. The claim follows. □

Let $s \in [0, 2]$. Denote by ℓ_2^s the set of real-valued sequences $a = (a_j)_{j\geq 0}$ such that

$$\|a\|_{\ell_2^s} := \left(\sum_{j=0}^{\infty}(2^{js}|a_j|)^2\right)^{\frac{1}{2}} < \infty$$

Furthermore, let $L^r_0(\ell_2^s)$ denote the space of all sequences $F = (f_j)_{j\geq 0} \subseteq L^r_0(\Omega)$ such that

$$\|F\|^r_{L^r(\ell_2^s)} := \int_\Omega \|F(x)\|^r_{\ell_2^s} dx < \infty.$$

By Theorems 5.1.2 and 5.6.3 of Bergh and Löfström (1976), the Banach spaces $B_s := L^r_0(\ell_2^s)$, for $0 \leq s \leq 2$, satisfy condition (i) of Lemma 50. Furthermore, the map

$$\mathscr{I} : \mathcal{H}^{s,r}(\Omega) \to B_s, \quad \mathscr{I} : u \mapsto \left(\sum_{\ell=1}^{\infty} \xi_j(\sqrt{\lambda_\ell})\mathcal{L}_\ell[u]\eta_\ell(\cdot)\right)_{j\geq 0}$$

satisfies, by Lemma 51, $\|u\|_{\mathcal{H}^{s,r}(\Omega)} \asymp \|u\|_{\mathscr{F}^s_{r,2}(\Omega)} = \|\mathscr{I}u\|_{B_s}$, and thus satisfies condition (ii) of Lemma 50. Finally, define the map

$$\mathscr{P} : B_s \to \mathcal{H}^{s,r}(\Omega), \quad \mathscr{P} : (f_j)_{j\geq 0} \mapsto \sum_{\ell=1}^{\infty}\sum_{j=0}^{\infty} \widetilde{\xi}_j(\sqrt{\lambda_\ell})\mathcal{L}_\ell[f_j]\eta_\ell,$$

where

$$\widetilde{\xi}_j = \xi_{j-1} + \xi_j + \xi_{j+1}, \quad j = 1, 2, \ldots$$

with the convention that $\xi_{-m} = 0$ for any $m > 0$. Notice that for all $u \in \mathcal{H}^{s,r}(\Omega)$, we have

$$\mathscr{P}\mathscr{I}u = \sum_{\ell=1}^{\infty}\sum_{j=0}^{\infty} \widetilde{\xi}_j(\sqrt{\lambda_\ell})\xi_j(\sqrt{\lambda_\ell})\mathcal{L}_\ell[u]\eta_\ell.$$

Since ξ_j has disjoint support from ξ_k for any $j, k \in \mathbb{N}$, $|j - k| \geq 2$, we have

$$\widetilde{\xi}_j(\sqrt{\lambda_\ell})\xi_j(\sqrt{\lambda_\ell}) = \sum_{k=0}^{\infty} \xi_k(\sqrt{\lambda_\ell})\xi_j(\sqrt{\lambda_\ell}) = \xi_j(\sqrt{\lambda_\ell}),$$

where we used the fact that $\{\xi_j\}_{j\geq 0}$ forms a partition of unity. By reapplying this property, we obtain

$$\mathscr{P}\mathscr{I}u = \sum_{\ell=1}^{\infty}\sum_{j=0}^{\infty} \xi_j(\sqrt{\lambda_\ell})\mathcal{L}_\ell[u]\eta_\ell = \sum_{\ell=1}^{\infty} \mathcal{L}_\ell[u]\eta_\ell = u.$$

It thus remains to show that $\mathscr{P}$ is a bounded linear operator. To do so, we will make use of the Hardy-Littlewood maximal function

$$Mf(x) = \sup_{B \in \mathcal{B}_x} \frac{1}{\mathcal{L}(B)} \int_B f(y)dy, \quad x \in \Omega,$$

for any $f \in L^1(\Omega)$, where $\mathcal{B}_x$ is the set of balls of the form $\{y \in \Omega : \|x - y\| < \delta\}$, $\delta > 0$. For our purposes, the utility of the maximal function lies in the fact that it induces a bounded operator from B_s into itself (cf. Theorem 5.6.6 of Grafakos (2009)): there exists a constant $C > 0$ depending on Ω, r such that for any $s \in [0, 2]$ and $(f_j)_{j\geq 0} \in B_s$, it holds that

$$\left\| \left(\sum_{j=0}^{\infty} (2^{js} |M(f_j)|)^2 \right)^{\frac{1}{2}} \right\|_{L^r(\Omega)} \leq C \left\| \left(\sum_{j=0}^{\infty} (2^{js} |f_j|)^2 \right)^{\frac{1}{2}} \right\|_{L^r(\Omega)}. \tag{5.74}$$

Let us now turn to bounding the operator norm of $\mathscr{P}$. Using Lemma 51, we have

$$\begin{aligned} \|\mathscr{P}(f_j)_j\|_{\mathcal{H}^{s,r}(\Omega)} &\asymp \left\| \left(\sum_{j=0}^{\infty} \left(2^{js} \sum_{\ell=1}^{\infty} \sum_{k=0}^{\infty} \xi_j(\sqrt{\lambda_\ell}) \widetilde{\xi}_k(\sqrt{\lambda_\ell}) \mathcal{L}[f_k]\eta_\ell \right)^2 \right)^{\frac{1}{2}} \right\|_{L^r(\Omega)} \\ &= \left\| \left(\sum_{j=0}^{\infty} \left(2^{js} \sum_{\ell=1}^{\infty} \sum_{k=j-2}^{j+2} \xi_j(\sqrt{\lambda_\ell}) \widetilde{\xi}_k(\sqrt{\lambda_\ell}) \mathcal{L}[f_k]\eta_\ell \right)^2 \right)^{\frac{1}{2}} \right\|_{L^r(\Omega)}, \end{aligned} \tag{5.75}$$

where we used the fact that ξ_j has disjoint support from $\widetilde{\xi}_k$ whenever $|j - k| \geq 3$. For any $j, k \geq 0$, let Λ_{jk} be the operator defined by

$$\Lambda_{jk} f = \sum_{\ell=1}^{\infty} \xi_j(\sqrt{\lambda_\ell}) \widetilde{\xi}_k(\sqrt{\lambda_\ell}) \mathcal{L}[f]\eta_\ell,$$

for any $f \in L^r_N(\Omega)$. In order to bound the right-hand side of equation (5.75), we will relate the operator Λ_{jk} to the maximal function M, in the following Lemma. This result is largely inspired by Georgiadis and Kyriazis (2023, Eq. (5.9)).

Lemma 52. *There exists a constant $C > 0$ such that for all $s \in [0, 2]$, $j \geq 0$, $f \in L^r_N(\Omega)$, and $x \in \Omega$,*

$$|\Lambda_{jk} f(x)| \leq CMf(x).$$

Before proving the Lemma, let us show how it implies the claim. Write $f_{-1} = f_{-2} = 0$. Continuing from equation (5.75), we obtain from Lemma 52 that

$$\|\mathscr{P}(f_j)_j\|_{\mathcal{H}^{s,r}(\Omega)} \lesssim \left\| \left(\sum_{j=0}^{\infty} \left(2^{js} \sum_{k=j-2}^{j+2} Mf_k \right)^2 \right)^{\frac{1}{2}} \right\|_{L^r(\Omega)}$$

$$
\begin{aligned}
&\lesssim \left\| \left(\sum_{j=0}^{\infty} \left(2^{js} M f_j\right)^2 \right)^{\frac{1}{2}} \right\|_{L^r(\Omega)} \\
&\lesssim \|(f_j)_j\|_{B_s},
\end{aligned}
$$

where the final inequality follows from equation (5.74). This proves the boundedness of the operator $\mathscr{P}$, and it remains to prove Lemma 52, which we shall do next.

By Theorem 3.1 of Kerkyacharian and Petrushev (2015) and the definition of the functions ξ_j, the operator Λ_{jk} is an integral operator,

$$
\Lambda_{jk} f(x) = \int_{\Omega} \Gamma_{jk}(x,y) f(y) dy, \quad f \in L^r(\Omega), x \in \Omega,
$$

where the kernel Γ_{jk} is real-valued and enjoys the bound

$$
|\Gamma_{jk}(x,y)| \lesssim_a \mathcal{L}\big(B(x, 2^{-j})\big)^{-1} (1 + 2^j \|x-y\|)^{-a},
$$

for any a sufficiently large, where $B(x,\delta) = \{y \in \Omega : \|x-y\| < \delta\}$. Note that the implicit constant above is independent of j,k. Under condition **(C1)**, we have $\mathcal{L}\big(B(x,2^{-j})\big)^{-1} \lesssim 2^{jd}$. Letting $D = \mathrm{diam}(\Omega)$, we thus have uniformly in $x \in \Omega$,

$$
\begin{aligned}
|\Lambda_{jk} f(x)| &= \left| \int_{\Omega} \Gamma_{jk}(x,y) f(y) dy \right| \\
&\leq \sum_{k=0}^{j-1} \int_{\{y \in \Omega : 2^{k-j} \leq \frac{\|x-y\|}{D} \leq 2^{k-j+1}\}} 2^{jd} (1 + 2^j \|x-y\|)^{-a} |f(y)| dy \\
&\leq \sum_{k=0}^{j-1} \int_{\{y \in \Omega : \frac{\|x-y\|}{D} \leq 2^{k-j+1}\}} 2^{jd-ka} |f(y)| dy \\
&\lesssim \sum_{k=0}^{j-1} 2^{jd-ka+d(k-j+1)} M f(x) \\
&\lesssim M f(x) \sum_{k=0}^{\infty} 2^{-k(a-d)}.
\end{aligned}
$$

Choosing $a > d$, we obtain the conclusion of Lemma 52, and hence of Proposition 32 in the regime $s \in [0,2]$, $r \geq 2$.

It remains to prove Proposition 32 in the regime $s \geq 0$ when $r = 2$. By Dunlop et al. (2020, Theorem 7.1), we already know that the claim holds for integer values of s. We will again use an interpolation argument to deduce the claim for non-integer values of s. Indeed, as noted by Dunlop et al. (2020), for $s_1 \in \mathbb{N}$, and any $0 \leq s \leq s_1$, the space $\mathcal{H}^s(\Omega)$ can be written as the following real interpolation space (Bergh and Löfström, 1976):

$$
\mathcal{H}^s(\Omega) = \big(L_0^2(\Omega), \mathcal{H}^{s_1}(\Omega)\big)_{s/s_1, 2}.
$$

Theorem 7.1 of Dunlop et al. (2020) thus implies

$$\mathcal{H}^s(\Omega) = \big(L_0^2(\Omega), H_N^{s_1}(\Omega)\big)_{s/s_1,2}.$$

The right-hand side of the above display is equal to $H_N^s(\Omega)$ by Löfström (1992), and the claim follows by taking s_1 arbitrarily large. □

5.I.2 Density Estimation under the $\mathcal{H}^{s,r}(\Omega)$ Norms

Our aim in this subsection is to prove the following result, which is our main technical tool for deriving Theorem 21. Let $p_{L_n}(x) := \mathbb{E}[\widetilde{p}_n(x)]$ for all $x \in \Omega$.

Proposition 33. Let $r \geq 2$, $M, s, c > 0$, and assume that $p \in \mathcal{C}_N^s(\Omega; M)$. Assume that either $s \leq 2$ or $r = 2$. Assume further that $L_n = cn^a$ for some $0 < a < 1$. Then, for any given $-\infty < t < s$, there exists a constant $C > 0$ depending on $\Omega, M, d, t, s, r, c, a$ such that

$$\|p_{L_n} - p\|_{\mathcal{H}^{t,r}(\Omega)}^r \leq CL_n^{-\frac{r(s-t)}{d}},$$

and, if we further assume $t \geq 0$,

$$\mathbb{E}\|\widetilde{p}_n - p_{L_n}\|_{\mathcal{H}^{t,r}(\Omega)}^r \leq Cn^{-\frac{r}{2}} L_n^{r\left(\frac{t}{d}+\frac{1}{2}\right)}.$$

In particular, if $L_n^{1/d} \asymp n^{\frac{1}{d+2s}}$, then for $t \geq 0$,

$$\mathbb{E}\|\widetilde{p}_n - p\|_{\mathcal{H}^{t,r}(\Omega)}^r \lesssim n^{-\frac{r(s-t)}{2s+d}}.$$

5.I.2.1 Proof of Proposition 33

We begin by bounding the variance term, assuming $t \geq 0$. Recall that $L_n \asymp n^a$ with $0 < a < 1$, and define $r_0 = 2/(1-a)$. We begin by proving the claim when $r \geq r_0$.

We wish to bound the quantity

$$V_n = \mathbb{E}\big\|(-\Delta)^{t/2}[\widetilde{p}_n - p_{L_n}]\big\|_{L^r(\Omega)}^r = \left\|\sum_{\ell=1}^{L_n} \mathcal{L}_\ell[\widetilde{p}_n - p_{L_n}]\omega_\ell \lambda_\ell^{t/2}\eta_\ell\right\|_{L^r(\Omega)}^r.$$

Notice that

$$\mathcal{L}_\ell[\widetilde{p}_n - p_{L_n}] = \widehat{\alpha}_\ell - \alpha_\ell = \frac{1}{n}\sum_{i=1}^n (\eta_\ell(X_i) - \mathbb{E}[\eta_\ell(X_i)]),$$

so that

$$V_n = \int_\Omega \left|\frac{1}{n}\sum_{i=1}^n U_{n,i}(x)\right|^r dx, \quad \text{with } U_{n,i}(x) = \sum_{\ell=1}^{L_n} \omega_\ell \lambda_\ell^{t/2}\eta_\ell(x)\big(\eta_\ell(X_i) - \mathbb{E}[\eta_\ell(X_i)]\big).$$

By Rosenthal's inequalities (Rosenthal, 1970, 1972), we deduce that

$$V_n \lesssim n^{-r/2} \int_\Omega \mathbb{E}[|U_{n,1}(x)|^2]^{r/2} dx + n^{1-r} \int_\Omega \mathbb{E}|U_{n,1}(x)|^r dx. \tag{5.76}$$

We will provide a somewhat crude bound on $\mathbb{E}|U_{n,1}(x)|^r$, followed by a sharp bound on $\mathbb{E}|U_{n,1}(x)|^2$. The density p is Hölder-smooth over Ω, and is in particular bounded. We thus have for any $x \in \Omega$,

$$\begin{aligned}
\mathbb{E}|U_{n,1}(x)|^r &\lesssim \left\| \sum_{\ell=1}^{L_n} \omega_\ell \lambda_\ell^{t/2} \eta_\ell(x) \eta_\ell \right\|_{L^r(\Omega)}^r \\
&\leq \lambda_{L_n}^{rt/2} \sup_{y \in \Omega} \left(\sum_{\ell=1}^{L_n} |\eta_\ell(x)||\eta_\ell(y)| \right)^r \leq \lambda_{L_n}^{rt/2} \left(e_{L_n}(x,x) e_{L_n}(y,y) \right)^{\frac{r}{2}},
\end{aligned}$$

where we define the *spectral function* of the Neumann Laplacian by

$$e_L(x,y) = \sum_{\ell=1}^{L} \eta_\ell(x) \eta_\ell(y), \quad x, y \in \Omega,\ L \geq 1.$$

We will make use of the following bound on the spectral function.

Lemma 53 (Hörmander (2007), Theorem 17.5.3). *There exists a constant $C > 0$ such that for all $L \geq 1$,*

$$\|e_L\|_{L^\infty(\Omega \times \Omega)} \leq C \lambda_L^{\frac{d}{2}}.$$

In order to bound the eigenvalue appearing on the right-hand side of the above Lemma, we make use of Weyl's Law for the Neumann Laplacian (see for instance Dunlop et al. (2020), Lemma 7.10, and references therein).

Lemma 54 (Weyl's Law). *There exists a constant $c > 0$ depending only on Ω such that*

$$\ell^{2/d}/c \leq \lambda_\ell \leq c \ell^{2/d}, \quad \ell = 1, 2, \dots$$

By Lemmas 53–54, we obtain

$$\mathbb{E}|U_{n,1}(x)|^r \lesssim \lambda_{L_n}^{\frac{rt}{2} + \frac{dr}{2}} \asymp L_n^{r\left(\frac{t}{d}+1\right)}. \tag{5.77}$$

Using Plancherel's identity, a sharper bound in available in the quadratic case:

$$\begin{aligned}
\mathbb{E}|U_{n,1}(x)|^2 &\lesssim \left\| \sum_{\ell=1}^{L_n} \omega_\ell \lambda_\ell^{t/2} \eta_\ell(x) \eta_\ell \right\|_{L^2(\Omega)}^2 \\
&= \sum_{\ell=1}^{L_n} \omega_\ell^2 \lambda_\ell^t \eta_\ell^2(x)
\end{aligned}$$

$$\begin{aligned}&\lesssim \lambda_{L_n}^t e_{L_n}(x,x)\\ &\lesssim \lambda_{L_n}^{t+\frac{d}{2}} \asymp L_n^{\frac{2t}{d}+1}. \end{aligned} \tag{5.78}$$

Combining equation (5.76) with the bounds (5.77) and (5.78), we have thus shown:

$$V_n \lesssim n^{-\frac{r}{2}} L_n^{r\left(\frac{t}{d}+\frac{1}{2}\right)} + n^{1-r} L_n^{r\left(\frac{t}{d}+1\right)}.$$

Since $r \geq r_0$, the second term is of lower order than the first, and we obtain the claimed bound

$$V_n \lesssim n^{-\frac{r}{2}} L_n^{r\left(\frac{t}{d}+\frac{1}{2}\right)}.$$

If we instead have $r < r_0$, then Jensen's inequality and the above bound imply

$$\begin{aligned}\mathbb{E}\big\|(-\Delta)^{t/2}[\widetilde{p}_n - p_{L_n}]\big\|_{L^r(\Omega)}^r &\leq \mathbb{E}\big\|(-\Delta)^{t/2}[\widetilde{p}_n - p_{L_n}]\big\|_{L^{r_0}(\Omega)}^r\\ &\leq \left(\mathbb{E}\big\|(-\Delta)^{t/2}[\widetilde{p}_n - p_{L_n}]\big\|_{L^{r_0}(\Omega)}^{r_0}\right)^{\frac{r}{r_0}}\\ &\lesssim \left(n^{-\frac{r_0}{2}} L_n^{r_0\left(\frac{t}{d}+\frac{1}{2}\right)}\right)^{\frac{r}{r_0}}\\ &= n^{-\frac{r}{2}} L_n^{r\left(\frac{t}{d}+\frac{1}{2}\right)}.\end{aligned}$$

This completes our bound of the fluctuations when $t \geq 0$.

We now turn to bounding the bias term, where we now allow t to be any real number. Our main technical tool will be the multiplier result in Lemma 48. Define the map

$$m(x) = |x|^{t-s}(1-\tau(|x|^2)), \quad x \in \mathbb{R}, \tag{5.79}$$

where we recall that τ is the function used to define the weights ω_j. Notice that $m(x) = 0$ for all $|x| \leq 1/2$. Furthermore, it is clear that $m \in \mathcal{C}^\infty(\mathbb{R}_+)$, and that m satisfies Mikhlin's condition (5.70). It is then also clear that the map $m(\cdot/\sqrt{\lambda_{L_n}})$ satisfies this condition.

Now, with the convention $\omega_\ell = 0$ for all $\ell \geq L_n + 1$,

$$(-\Delta)^{t/2}[p - p_{L_n}] = \sum_{\ell=1}^\infty (1-\omega_\ell)\lambda_\ell^{t/2}\alpha_\ell\eta_\ell = \lambda_{L_n}^{-\frac{s-t}{2}} \sum_{\ell=1}^\infty m(\sqrt{\lambda_\ell/\lambda_{L_n}})\lambda_\ell^{s/2}\alpha_\ell\eta_\ell,$$

thus, applying Lemma 48 to the multiplier $m(\cdot/\sqrt{\lambda_{L_n}})$, we obtain

$$\begin{aligned}\big\|(-\Delta)^{t/2}[p_{L_n} - p]\big\|_{L^r(\Omega)} &= \left\|\lambda_{L_n}^{-\frac{s-t}{2}} \sum_{\ell=1}^\infty m(\sqrt{\lambda_\ell/\lambda_{L_n}})\lambda_\ell^{s/2}\alpha_\ell\eta_\ell\right\|_{L^r(\Omega)}\\ &\lesssim \lambda_{L_n}^{-\frac{s-t}{2}} \left\|\sum_{\ell=1}^\infty \lambda_\ell^{s/2}\alpha_\ell\eta_\ell\right\|_{L^r(\Omega)}\\ &\lesssim {L_n}^{-\frac{s-t}{d}} \|p\|_{\mathcal{H}^{s,r}(\Omega)},\end{aligned}$$

where we again used Weyl's law. It thus suffices to show that $\|p\|_{\mathcal{H}^{s,r}(\Omega)}$ is finite. To this end, let r_1 be defined as r if $s - 1/r$ is not an odd integer, and otherwise define r_1 as $r + \delta$ for any small enough $\delta < 1$. Then, by Proposition 32 (with $s \leq 2$ or $r = r_1 = 2$) and the definitions of the spaces $\mathcal{C}^s_N(\Omega)$ and $H^{s,r}_N(\Omega)$, we have

$$\|p\|_{\mathcal{H}^{s,r}(\Omega)} \leq \|p\|_{\mathcal{H}^{s,r_1}(\Omega)} \lesssim \|p\|_{H^{s,r_1}(\Omega)} \leq \|p\|_{\mathcal{C}^s(\Omega)} \leq M.$$

The claim thus follows □

Remark 4. Suppose that instead of the estimator

$$\widetilde{p}_n = \sum_{\ell=1}^{L_n} \omega_\ell \widehat{\alpha}_\ell \eta_\ell$$

we had used the traditional truncated series estimator

$$\bar{p}_n = \sum_{\ell=1}^{L_n} \widehat{\alpha}_\ell \eta_\ell,$$

which corresponds to choosing the nonsmooth function $\tau(x) = I(|x| < 1)$ in the definition of the weights ω_ℓ. This choice would prevent the function m in equation (5.79) from satisfying the conditions of Lemma 48. In fact, if one were to replace Ω by $\mathbb{T}^d$, then the eigenvalues λ_ℓ would be of the form $\|2\pi\xi_\ell\|^2$ for some enumeration $\xi_1, \xi_2, \dots$ of $\mathbb{Z}^d_*$. In this case, we have for all $\ell \geq 1$,

$$1 - \tau(\lambda_\ell/\lambda_{L_n}) = I(\|\xi_\ell\| \leq \|\xi_{L_n}\|).$$

Viewed as a function of ξ_ℓ, the right-hand side is the indicator function of a ball, which is well-known not to be an $L^r(\mathbb{T}^d)$ Fourier multiplier for $r > 2$ and $d > 1$ (Fefferman, 1971), thus the expression m in (5.79) is also not an $L^r(\mathbb{T}^d)$-multiplier in this case. This suggests that our current proof technique cannot be used for the traditional series estimator.

We now supplement Proposition 33 with a simple variance bound for negative values of t, but now focusing on the case $r = 2$.

Proposition 34. Let $M, s > 0$, and assume that $p \in \mathcal{C}^s_N(\Omega; M)$. Then, for any given $t < 0$, there exists a constant $C > 0$ depending on Ω, M, d, t, s such that

$$\mathbb{E}\|\widetilde{p}_n - p_{L_n}\|^2_{\mathcal{H}^t(\Omega)} \lesssim \frac{1}{n} \begin{cases} L_n^{\frac{2t}{d}+1}, & 2|t| < d, \\ \log(L_n), & 2|t| = d, \\ 1, & 2|t| > d. \end{cases}$$

In particular, if $L_n^{1/d} \asymp n^{\frac{1}{d+2s}}$, then for $t \geq 0$,

$$\mathbb{E}\|\widetilde{p}_n - p\|^2_{\mathcal{H}^t(\Omega)} \lesssim \begin{cases} n^{-\frac{2(s-t)}{2s+d}}, & 2|t| < d \\ \log n/n, & 2|t| = d \cdot \\ 1/n, & 2|t| > d \end{cases}$$

Proof of Proposition 34. Notice that

$$\mathbb{E}\|\widetilde{p}_n - p_{L_n}\|^2_{\mathcal{H}^t(\Omega)} = \mathbb{E}\left[\sum_{\ell=1}^{L_n} \lambda_\ell^t \omega_\ell^2 (\widehat{\alpha}_\ell - \alpha_\ell)^2\right] \leq \sum_{\ell=1}^{L_n} \lambda_\ell^t \operatorname{Var}[\widehat{\alpha}_\ell] = \sum_{\ell=1}^{L_n} \lambda_\ell^t \frac{\operatorname{Var}[\eta_\ell(X)]}{n}.$$

Since p is bounded from above by M over Ω, we have

$$\operatorname{Var}[\eta_\ell(X)] \leq \mathbb{E}[\eta_\ell^2(X)] \leq M\|\eta_\ell\|^2_{L^2(\Omega)} = M,$$

thus, together with Weyl's Law, we have

$$\mathbb{E}\|\widetilde{p}_n - p_{L_n}\|^2_{\mathcal{H}^{-t}(\Omega)} \leq \frac{M}{n}\sum_{\ell=1}^{L_n} \ell^{-2|t|/d}.$$

The claim follows from here. □

5.I.3 Regularity of the Density Estimator

With Proposition 33 in hand, we can prove the following result, which will allow us to invoke condition **(C2)** directly on the density estimators.

Lemma 55. *Let $M, \gamma, s, c > 0$, and assume that $p \in \mathcal{C}_N^s(\Omega; M, \gamma)$. Assume $L_n^{1/d} = cn^{\frac{1}{d+2s}}$. Then, there exist constants $C, \epsilon > 0$ depending on $\Omega, M, \gamma, d, s, c$ such that the following assertions hold on an event of probability at least $1 - C/n^2$:*

(i) $\widetilde{p}_n \geq 1/C$ *over* Ω*. In particular,* $\widetilde{p}_n = \widehat{p}_n$*.*

(ii) $\|\widetilde{p}_n\|_{\mathcal{C}^\epsilon(\Omega)} \leq C$*.*

Proof of Lemma 55. Let $t = (s/2) \wedge 1$ and $\epsilon = t/2$. By a Sobolev embedding (cf. Triebel (1995), Theorem 4.6.1), we have for all $r \geq r_0 := 2d/\epsilon$,

$$\|\widetilde{p}_n - p\|_{\mathcal{C}^\epsilon(\Omega)} \lesssim \|\widetilde{p}_n - p\|_{H^{t,r}(\Omega)} \asymp \|\widetilde{p}_n - p\|_{\mathcal{H}^{t,r}(\Omega)},$$

where the final order assessment follows from Proposition 32. By Proposition 33, we thus obtain

$$\mathbb{E}\|\widetilde{p}_n - p\|^r_{\mathcal{C}^\epsilon(\Omega)} \lesssim n^{-\frac{r(s-t)}{2s+d}}.$$

Let $u = \big(\|p^{-1}\|^{-1}_{L^\infty(\Omega)} \wedge \|p\|_{\mathcal{C}^\epsilon(\Omega)}\big)/2$. Then, by Markov's inequality, we have

$$\mathbb{P}\Big(\|\widetilde{p}_n - p\|_{\mathcal{C}^\epsilon(\Omega)} \geq u\Big) \leq \frac{\mathbb{E}\|\widetilde{p}_n - p\|^r_{\mathcal{C}^\epsilon(\Omega)}}{u^r} \lesssim \frac{n^{-\frac{r(s-t)}{2s+d}}}{u^r}.$$

Since $t < s$, we may choose r large enough such that $\frac{r(s-t)}{2s+d} \geq 2$. Thus we readily deduce that for a large enough constant $C > 0$, we have with probability at least $1 - C/n^2$ that

$$\|\widetilde{p}_n - p\|_{\mathcal{C}^\epsilon(\Omega)} \leq u.$$

Over the above high-probability event, we have on the one hand

$$\inf_{x\in\Omega}\widetilde{p}_n(x) \geq \inf_{x\in\Omega}p(x) - \|\widetilde{p}_n - p\|_{L^\infty(\Omega)} \geq \inf_{x\in\Omega}p(x) - \|\widetilde{p}_n - p\|_{\mathcal{C}^\epsilon(\Omega)} \geq \inf_{x\in\Omega}p(x)/2,$$

from which part (i) of Lemma 55 follows. On the other hand,

$$\|\widetilde{p}_n\|_{\mathcal{C}^\epsilon(\Omega)} \leq \|p\|_{\mathcal{C}^\epsilon(\Omega)} + \|\widetilde{p}_n - p\|_{\mathcal{C}^\epsilon(\Omega)} \leq 2\|p\|_{\mathcal{C}^\epsilon(\Omega)},$$

from which part (ii) of Lemma 55 follows. □

5.I.4 Convergence Rate under the Wasserstein Distance

With the help of Proposition 33 we can obtain a bound on the risk of $\widehat{P}_n$ in Wasserstein distance.

Lemma 56. *Let $M,\gamma,s>0$, and assume that $p\in\mathcal{C}^s_N(\Omega;M,\gamma)$. Assume $L_n^{1/d}=cn^{\frac{1}{d+2s}}$. Then, there exists a constant $C>0$ depending on Ω,M,γ,d,s,r,c such that*

$$\mathbb{E}W_2^2(\widehat{P}_n,P) \lesssim \begin{cases} n^{-\frac{2(s+1)}{2s+d}}, & d>2 \\ \log n/n, & d=2 \\ 1/n, & d=1. \end{cases}$$

Proof of Lemma 56. Let A_n be the event over which the two assertions of Lemma 55 hold. By the bound (5.9) due to Peyre (2018) (in its form stated in Theorem 5.34 of Santambrogio (2015)), it holds that

$$\mathbb{E}W_2^r(\widehat{P}_n,P) \lesssim \mathbb{E}\big[W_2^r(\widehat{P}_n,P)I(A_n)\big] + n^{-2} \lesssim \mathbb{E}\big[\|\widetilde{p}_n - p\|^r_{\mathcal{H}^{-1}(\Omega)}\big] + n^{-2}.$$

The claim thus follows from Proposition 34. □

5.I.5 Proof of Theorem 21

Using Lemmas 55 and 56, Theorem 21(i) follows by the same argument as Theorem 19(i), and Theorem 21(ii) follows by the same argument as Theorems 20 and 22. In the latter case, one replaces the application of Caffarelli's regularity theory (Theorem 4) by an application of condition **(C2)**. We omit further details for brevity. □

Chapter 6

Central Limit Theorems for Smooth Optimal Transport Maps

6.1 Introduction

Our aim is now to build upon the results of the previous chapter, to address the problem of *inference* for Brenier maps. Specifically, our goal is to initiate the study of limit laws for Brenier maps between absolutely continuous distributions in general dimension.

We will focus our attention throughout on probability measures supported on the d-dimensional flat torus $\mathbb{T}^d$. Let P and Q denote two absolutely continuous probability measures with respect to the uniform law on $\mathbb{T}^d$, with respective densities p and q, and let $T_0 = \nabla\varphi_0$ be the Brenier map pushing forward P onto Q, with respect to a Brenier potential φ_0. We will always assume in this chapter that $p, q, 1/p$, and $1/q$ are bounded functions over $\mathbb{T}^d$. As we recall below, these conditions imply that T_0 admits a unique representative which is continuous over $\mathbb{T}^d$ (Caffarelli, 1992), and we always assume that T_0 is taken to be this representative. Likewise, the potential φ_0 admits a representative which is continuously differentiable and uniquely defined up to an additive constant, and we will always assume that φ_0 is taken to be this representative, with the additive constant chosen such that φ_0 has mean zero over the unit hypercube. With this convention, T_0 and φ_0 are uniquely defined, and can both be evaluated pointwise without any ambiguity.

We will be primarily interested in the one-sample problem, in which only the measure Q is sampled from. As discussed in Remark 5, our techniques can immediately be extended to the case where P is also sampled from, however we do not carry out this extension in order to keep our exposition concise. Let $Q_n = (1/n)\sum_{i=1}^n \delta_{Y_i}$ be an empirical measure comprised of i.i.d. random variables $Y_1, \dots, Y_n$ with common law Q. Our main object of interest is the optimal transport map which pushes forward p onto the kernel density estimator of q, which we recall is defined as

$$\widehat{q}_n = Q_n \star K_{h_n} = \int_{\mathbb{R}^d} K_{h_n}(\cdot - y)dQ_n(y),$$

where, as in the previous chapter, $K \in \mathcal{C}^\infty(\mathbb{R}^d)$ is a smooth mollifier which integrates to unity, $K_{h_n}(\cdot) = K(\cdot/h_n)/h_n^d$ for some nonnegative sequence $h_n \downarrow 0$, and where the integration in the above display is to be interpreted by extending Q_n to a measure on $\mathbb{R}^d$ via $\mathbb{Z}^d$-periodicity. The kernel K will typically be permitted to take on negative values, thus $\widehat{q}_n$ does not necessarily define a probability density. When it does, we define $\widehat{Q}_n$ to be the probability law with density $\widehat{q}_n$[1]. Furthermore, over the event that $\widehat{q}_n$ and $1/\widehat{q}_n$ are nonnegative bounded functions on $\mathbb{T}^d$, we define $\widehat{T}_n$ to be the unique continuous optimal transport map pushing P forward onto $\widehat{Q}_n$, and $\widehat{\varphi}_n$ to be the unique continuously differentiable Brenier potential, admitting mean zero over the unit hypercube, such that $\widehat{T}_n = \nabla\widehat{\varphi}_n$. Over the complement of this event, $\widehat{T}_n$ can be defined as any continuous vector field from $\mathbb{T}^d$ into itself, without changing any of our conclusions; for instance, one may take $\widehat{T}_n = \mathrm{Id}$.

When K is taken to be a Gaussian kernel, the density estimator $\widehat{q}_n$ is simply the regularization of the empirical measure Q_n with respect to the heat kernel at time h_n^2. This form of regularization has received a great deal of recent interest in the optimal transport literature, ever since it was used by Ambrosio, Glaudo, and Trevisan (2019); Ambrosio, Stra, and Trevisan (2019), building upon a conjecture of Caracciolo et al. (2014), to derive exact asymptotics for the quadratic optimal matching problem. Unlike these works, however, our aim here is not to treat $\widehat{T}_n$ as an approximation of the Brenier map T_n pushing P forward onto Q_n. Indeed, our results will typically require h_n to vanish at too slow of a rate for such an approximation to be meaningful. We instead view $\widehat{T}_n$ as the object of interest, motivated by the fact that it provides a better approximation of T_0 than the empirical map T_n, when the underlying densities are smooth.

A close variant of the estimator $\widehat{T}_n$ has already appeared in Chapter 5 (see also the works of Gunsilius (2022) and Deb, Ghosal, and Sen (2021)), where our main goal was to derive its expected $L^2(\mathbb{T}^d)$ convergence rate; cf. Theorem 20. While this result provides an essentially sharp characterization of the L^2 convergence rate of $\widehat{T}_n$, it does not offer any insight into its pointwise or uniform behaviour. The main result of this chapter is to derive the pointwise limiting distribution of $\widehat{T}_n$, in the regime $d \geq 3$.

Theorem (Informal)**.** Let $s > 2$, $d \geq 3$, and assume $p, q \in \mathcal{C}_+^s(\mathbb{T}^d)$. Then, there exists a sequence $h_n = o(h_n^*)$ such that for all $x \in \mathbb{T}^d$, there exists a positive definite $d \times d$ matrix $\Sigma(x)$ such that, as $n \to \infty$,

$$\sqrt{nh_n^{d-2}}\big(\widehat{T}_n(x) - T_0(x)\big) \xrightarrow{w} N(0, \Sigma(x)). \tag{6.1}$$

The limiting covariance $\Sigma(x)$ can be made completely explicit, and will be described in

[1] Unlike the previous chapter, where we denoted the kernel density estimator by $\widetilde{q}_n$, and by $\widehat{q}_n^{(\mathrm{ker})}$ its positive part (normalized to be a density), our results in this section will be more concisely stated for the kernel density estimator itself, which we write as $\widehat{q}_n \equiv \widetilde{q}_n$

Theorem 24 below. Furthermore, the result in fact holds for a wide range of sequences h_n which are of lower order than h_n^*. As we shall see, the latter condition is crucial for the centering constant in equation (6.1) to be the population quantity $T_0(x)$ itself.

To the best of our knowledge, this result is the first central limit theorem for optimal transport maps between absolutely continuous distributions in dimension greater than one. The one-dimensional counterpart of this result was established by Ponnoprat, Okano, and Imaizumi (2024), in which case the appropriate scaling sequence is simply $\sqrt{n}$ rather than $\sqrt{nh_n^{d-2}}$, as could have heuristically been anticipated from Theorem 20. Their work makes use of the representation of univariate optimal transport maps in terms of quantile functions, which of course cannot be exploited in general dimension. We also note that limit laws for related problems have appeared in recent literature. In the case where P and Q are discrete distributions, Klatt, Munk, and Zemel (2022) have derived central limit theorems for optimal transport couplings, while del Barrio, González-Sanz, and Loubes (2024) and Sadhu, Goldfeld, and Kato (2023) considered the case where only one of P or Q is discrete. Furthermore, Harchaoui, Liu, and Pal (2020), Gunsilius and Xu (2021), González-Sanz, Loubes, and Niles-Weed (2022), Goldfeld et al. (2024), and González-Sanz and Hundrieser (2023) derived limit laws for entropically regularized variants of the optimal transport map with a fixed regularization parameter. Let us emphasize that in each of these past works, the limit laws hold in a weak sense, and their scaling is of the parametric order $\sqrt{n}$, which is a reflection of the fact that the collection of optimal transport potentials arising in these problems forms a Donsker class. This property fails to hold in our setting, as can be anticipated from the fact that our pointwise central limit theorem exhibits a nonparametric scaling $\sqrt{nh_n^{d-2}}$. Furthermore, one cannot hope for a weak limit law to hold under our assumptions: we will show in Theorem 25 below that there is no scaling of the process $\widehat{T}_n - T_0$ which converges to a non-degenerate limit in $L^2(\mathbb{T}^d)$, for a sensible range of values h_n. Such a result is akin to the failure of weak limit laws for the kernel density estimator itself (Nishiyama, 2011; Stupfler, 2014, 2016).

Let us provide a heuristic summary of our proof strategy. Our starting point is Caffarelli's interior regularity theory for optimal transport maps, which we recalled in Theorem 3 of Chapter 1. In their strongest form, these results imply that, under the smoothness condition $p, q \in \mathcal{C}_+^s(\mathbb{T}^d)$, the Brenier potential φ_0 lies in $\mathcal{C}^2_{\mathrm{loc}}(\mathbb{R}^d)$. In fact, we will see that with probability tending to one, the *estimator* $\widehat{\varphi}_n$ itself lies in $\mathcal{C}^2_{\mathrm{loc}}(\mathbb{R}^d)$, with Hölder norm over any fixed compact set which is uniformly bounded in n. Over this high-probability event, the pushforward conditions $\widehat{T}_{n\#}P = \widehat{Q}_n$ and $T_{0\#}P = Q$ are equivalent to the solvability of the following *Monge-Ampère equations*

$$\det(\nabla^2\widehat{\varphi}_n) = \frac{p}{\widehat{q}_n(\nabla\widehat{\varphi}_n)}, \quad \text{and} \quad \det(\nabla^2\varphi_0) = \frac{p}{q(\nabla\varphi_0)}, \quad \text{over } \mathbb{R}^d,$$

in the classical sense. It is well-known that these equations formally linearize to second-order uniformly elliptic partial differential equations (Villani, 2003). By leveraging this fact, we use a Taylor expansion argument to show that

$$\widehat{T}_n = T_0 + \nabla u_n + R_n, \tag{6.2}$$

where R_n is a Taylor remainder which vanishes at a fast rate in $L^\infty(\mathbb{T}^d)$, and where u_n is the unique mean-zero solution to the partial differential equation

$$Lu_n = -\mathrm{div}(q\nabla u_n(\nabla\varphi_0^*)) = \widehat{q}_n - q, \quad \text{over } \mathbb{T}^d. \tag{6.3}$$

Here, φ_0^* denotes the Legendre-Fenchel transform of φ_0. Let us briefly comment on equation (6.3). In the special case where p and q are both equal to the uniform law on $\mathbb{T}^d$, the above is simply the periodic Poisson equation, which has appeared in similar linearization strategies in the works of Ambrosio, Glaudo, and Trevisan (2019), Ambrosio, Stra, and Trevisan (2019), Ambrosio and Glaudo (2019), Goldman and Huesmann (2022), Clozeau and Mattesini (2023), and references therein, in the context of deriving exact asymptotics for various optimal matching problems. When p and q are equal but not necessarily uniform, the differential operator above becomes the Witten Laplacian $-\mathrm{div}(q\nabla\cdot)$, as noted by Greengard et al. (2022). For general densities p and q, the operator L is not strictly-speaking of elliptic type, but with enough regularity assumptions on φ_0, we will show that it is closely connected to a self-adjoint uniformly elliptic operator.

In view of the formal expansion (6.2), the problem of deriving a central limit theorem for $\widehat{T}_n$ reduces to the problem of deriving a pointwise central limit theorem for the gradient of the solution to the PDE (6.3). While L^2 rates for estimating coefficients of elliptic PDEs have been studied in the literature (cf. Nickl, Van De Geer, and Wang (2020), Giordano and Nickl (2020), and references therein), we are not aware of existing pointwise convergence rates or limit laws for solutions to elliptic PDEs with a stochastic right-hand side. The bulk of our effort goes into this point. Letting Q_{h_n} be the probability distribution with density $q_{h_n} = \mathbb{E}[\widehat{q}_n(\cdot)]$, we begin by decomposing ∇u_n into the terms

$$\nabla u_n(x) = \nabla L^{-1}[\widehat{q}_n - q_{h_n}](x) + \nabla L^{-1}[q_{h_n} - q](x), \quad \text{for any given } x \in \mathbb{T}^d. \tag{6.4}$$

The first (stochastic) term on the right-hand side of the above display will characterize the fluctuations of $\widehat{T}_n(x)$ around its mean, while the second (deterministic) term will characterize the bias of $\widehat{T}_n(x)$. We will show that the condition $h_n = o(h_n^*)$ is esentially sufficient for the deterministic term to be of negligible order; the heart of the matter lies in the stochastic term. To derive its asymptotic distribution, we appeal to a variant of the "coefficient freezing" method which is commonly used to derive a priori estimates for elliptic PDE. Specifically, we will argue that, in a neighborhood of the point $T_0(x)$, the solution to equation (6.3) is well-approximated by the solution to the equation

$$-q(T_0(x))\mathrm{div}(\nabla v_n(G_x)) = \widehat{q}_n - q_{h_n}, \quad \text{over } \mathbb{T}^d, \tag{6.5}$$

where G_x denotes the first-order Taylor approximation of $\nabla\varphi_0^*$ at the point $T_0(x)$. This approximation is useful because, unlike ∇u_n, the map ∇v_n can be expressed as the evaluation of a *convolution operator* at the smoothed empirical process. Indeed, we will show that there exists a map $\Theta : \mathbb{T}^d \to \mathbb{R}^d$ for which the following representation holds:

$$\nabla v_n(x) = \int_{\mathbb{T}^d} \Theta(y - T_0(x))d(\widehat{Q}_n - Q_{h_n})(y). \tag{6.6}$$

The map Θ is the gradient of a periodic Green's function associated to the differential operator appearing on the left-hand side of equation (6.5), which can be derived in closed form when working over the torus. By associativity of convolutions, we can further write

$$\nabla v_n(x) = \big[(\Theta \star K_{h_n}) \star (Q_n - Q)\big](T_0(x)).$$

The map in the above display is manifestly a kernel density estimator with respect to the vector-valued "kernel" $\Theta \star K_{h_n}$. Its limiting distribution can be derived using elementary means, and ultimately describes the limiting distribution of $\widehat{T}_n(x)$.

Though our main interest is in pointwise limit laws, as a byproduct of these results, we will also derive nonasymptotic pointwise rates of convergence for the estimator $\widehat{T}_n$. A great deal of recent work has analyzed L^2 rates of estimation for optimal transport maps (Hütter and Rigollet, 2021; Deb and Sen, 2021; Pooladian and Niles-Weed, 2021; Ghosal and Sen, 2022; Gunsilius, 2022; Divol, Niles-Weed, and Pooladian, 2022; Pooladian, Divol, and Niles-Weed, 2023), and qualitative uniform convergence results have been studied in the works of Chernozhukov et al. (2017); Panaretos and Zemel (2019); Hallin et al. (2021); De Lara, González-Sanz, and Loubes (2021); Ghosal and Sen (2022); Segers (2022), but our work is perhaps the first to provide near-optimal rates for pointwise estimation (albeit under the strong assumption that the underlying domain is the flat torus).

6.2 Main Results

In this section, we state our main results regarding the asymptotic behaviour of the Brenier map $\widehat{T}_n$. We begin by stating one of our key technical results, namely a quantitative linearization estimate for the Monge-Ampère equation. For the remainder of the chapter, we fix once and for all a real number $s > 2$ such that $s \notin \mathbb{N}$, which will typically represent the Hölder smoothness exponent of the densities p and q.

6.2.1 Linearization of the Monge-Ampère Equation

Define a constant β such that

$$0 < \beta < \min\{1, s-2\}. \tag{6.7}$$

Given $p, q \in \mathcal{C}_+^{2+\beta}(\mathbb{T}^d)$, it follows from Caffarelli's regularity theory (Theorem 3) that $\varphi_0, \varphi_0^* \in \mathcal{C}^{4+\beta}(\mathbb{T}^d)$. We may then define the operator

$$Lu = -\mathrm{div}(q \nabla u(\nabla \varphi_0^*)),$$

for all $u \in \mathcal{C}^2(\mathbb{T}^d)$. As we shall see in Section 6.3, it can be deduced from standard elliptic regularity theory that L is a bijection of $\mathcal{C}_0^{2+\beta}(\mathbb{T}^d)$ onto $\mathcal{C}_0^{\beta}(\mathbb{T}^d)$, with inverse denoted L^{-1}.

Let $\widehat{Q} \in \mathcal{P}_{\mathrm{ac}}(\mathbb{T}^d)$ be any distribution with density $\widehat{q}$ over $\mathbb{T}^d$, and assume that $\widehat{q} \in \mathcal{C}_+^{2+\beta}(\mathbb{T}^d)$. Let $\widehat{\varphi}$ be the unique Brenier potential lying in $\mathcal{C}^2(\mathbb{R}^d)$ whose gradient pushes forward P onto $\widehat{Q}$, and which satisfies $\int_{\mathcal{Q}} \widehat{\varphi} d\mathcal{L} = 0$, where we define once and for all

$$\mathcal{Q} := [0,1]^d.$$

The following result shows that, in a very strong sense, the deviations $\widehat{\varphi} - \varphi_0$ are well-approximated by the solution to a linear partial differential equation, whenever $\widehat{q}$ is in a Hölder neighborhood of q.

Theorem 23. *Let $d \geq 1$, and assume $p, q, \widehat{q} \in \mathcal{C}^{2+\beta}_+(\mathbb{T}^d)$. Then, there exists a constant $C = C(\omega_{2+\beta}(p, q, \widehat{q}), d, \beta) > 0$ such that*

$$\big\|(\widehat{\varphi} - \varphi_0) - L^{-1}[\widehat{q} - q]\big\|_{\mathcal{C}^{2+\beta}(\mathbb{T}^d)} \leq C\|\widehat{q} - q\|_{\mathcal{C}^{1+\beta}(\mathbb{T}^d)}\|\widehat{q} - q\|_{\mathcal{C}^{\beta}(\mathbb{T}^d)}. \tag{6.8}$$

Theorem 23 is proved in Section 6.3. In particular, this result implies that the approximation

$$\nabla\widehat{\varphi} - \nabla\varphi_0 \sim \nabla L^{-1}[\widehat{q} - q] \tag{6.9}$$

holds, up to an error which decays in uniform norm on the order $\|\widehat{q} - q\|_{\mathcal{C}^{1+\beta}(\mathbb{T}^d)}\|\widehat{q} - q\|_{\mathcal{C}^{\beta}(\mathbb{T}^d)}$. We believe it should be possible to replace the latter quantity by $\|\widehat{q} - q\|^2_{\mathcal{C}^{\beta}(\mathbb{T}^d)}$, but our current statement is sufficiently sharp for the applications which we have in mind. In Appendix B, we state a convergence rate for the kernel density estimator under Hölder norms, which will allow us to bound this approximation error when $\widehat{q}$ is taken to be the random density $\widehat{q}_n$.

Our proof of Theorem 23 relies on the following Hölder stability bound for Brenier potentials, which may be of independent interest: under the same conditions as Theorem 23, we show in Proposition 36 below that

$$\|\widehat{\varphi} - \varphi_0\|_{\mathcal{C}^{2+\beta}(\mathbb{T}^d)} \lesssim \|\widehat{q} - q\|_{\mathcal{C}^{\beta}(\mathbb{T}^d)}. \tag{6.10}$$

Equation (6.25) is closely related to a qualitative Sobolev stability result for the Monge-Ampère equation proven by De Philippis and Figalli (2013). We emphasize that their assumptions are weaker than ours, and in particular do not imply that the Brenier potentials are twice differentiable, which makes their analysis significantly more challenging. Building upon their result, Gunsilius (2022) derived a qualitative stability result for Brenier potentials under the $\mathcal{C}^{2+\beta}$ norm, but we are not aware of any quantiative bounds akin to the one above.

The proof of Theorem 23 appears in Section 6.3. Let us now show how this result can be used to analyze the estimator $\widehat{T}_n$.

6.2.2 Pointwise Convergence Rate

Before presenting central limit theorems, it will be fruitful to state a bound on the pointwise convergence rate of the estimator $\widehat{T}_n$, as this will make clear the scaling and centering that one may expect in the limit laws. We begin with some notation. Let $h_n^* = n^{-1/(d+2s)}$. For all $x \in \mathbb{T}^d$ and $h_n > 0$, set

$$T_{h_n}(x) = \mathbb{E}[\widehat{T}_n(x)], \quad \text{and} \quad q_{h_n}(x) = \mathbb{E}[\widehat{q}_n(x)].$$

Recall condition $\mathbf{K}(\boldsymbol{\alpha})$ on the kernel K defined in Chapter 5.

Proposition 35. Let $d \geq 3$, and $p, q \in \mathcal{C}^s_+(\mathbb{T}^d)$. Assume that the kernel K satisfies condition $\mathbf{K(s+1)}$, and that for some $c > 0$,

$$h_n = c \cdot n^{-a}, \quad \text{with } \frac{1}{d+4(s-1)} < a < \frac{1}{d+s+2}. \tag{6.11}$$

Then, for any $\epsilon > 0$, there exists a constant $C = C(\omega_s(p,q), K, s, c, a, \epsilon, d) > 0$ such that the following assertions hold.

1. (Bias) We have,

$$\big\|T_{h_n} - T_0\big\|_{L^\infty(\mathbb{T}^d)} \leq C h_n^{s+1-\epsilon}. \tag{6.12}$$

2. (Fluctuations) We have,

$$\sup_{x \in \mathbb{T}^d} \mathbb{E}\big\|\widehat{T}_n(x) - T_{h_n}(x)\big\| \leq C \frac{1}{\sqrt{n h_n^{d-2}}}. \tag{6.13}$$

The proof of Proposition 35 appears in Section 6.5. Condition (6.11) on the bandwidth h_n is never vacuous due to the assumption that $s > 2$. In particular, the bandwidth $h_n \asymp h_n^*$ satisfies this condition, and for this choice, the two bounds (6.12)–(6.13) are essentially on the same order. In this case, Proposition 35 implies

$$\sup_{x \in \mathbb{T}^d} \mathbb{E}\big\|\widehat{T}_n(x) - T_0(x)\big\| \lesssim n^{-\frac{s+1-\epsilon}{d+2s}}. \tag{6.14}$$

This result shows that the $L^2(\mathbb{T}^d)$ convergence rate of $\widehat{T}_n$, stated in Theorem 20, in fact holds in a pointwise sense, at the price of the exponent ϵ, which can be made arbitrarily small. Furthermore, by a simple modification of Theorem 6 of Hütter and Rigollet (2021), it can be seen that, in an information-theoretic sense, no other estimator can achieve a rate faster than $n^{-(s+1)/(2s+d)}$ for estimating $T_0(x)$ at a point $x \in \mathbb{T}^d$, uniformly over all measures P, Q satisfying the conditions of Proposition 35. In this sense, $\widehat{T}_n$ is a minimax optimal estimator, again, up to the arbitrarily small exponent ϵ. Though we conjecture that this exponent is superfluous, it cannot be easily removed with our current proof technique. Its presence is related to the fact that the Poisson equation $\Delta u = f$ on $\mathbb{T}^d$ is not necessarily solvable in the classical sense for $f \in L_0^\infty(\mathbb{T}^d)$, but admits a solution $u \in \mathcal{C}^{2+\epsilon}(\mathbb{T}^d)$ whenever $f \in \mathcal{C}_0^\epsilon(\mathbb{T}^d)$.

Proposition 35 is comparable to pointwise bounds for density estimation. It is well-known that, under weaker conditions than those of Proposition 35, it holds that for all $x \in \mathbb{T}^d$, and all dimensions $d \geq 1$ that (Giné and Nickl, 2016)

$$|q_{h_n}(x) - q(x)| \lesssim h_n^s, \quad \text{and} \quad \mathbb{E}|\widehat{q}_n(x) - q_{h_n}(x)| \lesssim \frac{1}{\sqrt{n h_n^d}}.$$

Both of these bounds are slower than those of Proposition 35 by a factor of h_n^{-1}, which is a reflection of the fact that optimal transport maps typically enjoy one degree of smoothness more than the corresponding densities. This can be inferred heuristically from equation (6.9). We nevertheless note that the optimal order of the bandwidth h_n which minimizes the sum of the above two terms is h_n^*, as in the case of Proposition 35.

Proposition 35 suggests that if the sequence of random variables

$$\sqrt{nh_n^{d-2}}(\widehat{T}_n(x) - T_0(x))$$

were to admit a non-degenerate limiting distribution, then h_n would have be taken of lower order than the optimal bandwidth h_n^*, a condition often referred to as *undersmoothing*. We derive limit laws under this condition, next.

6.2.3 Pointwise Central Limit Theorem

Our main result is the following.

Theorem 24. *Let $d \geq 3$, and $p, q \in \mathcal{C}_+^s(\mathbb{T}^d)$. Assume that condition* $\mathbf{K(s+1)}$ *holds, and*

$$h_n \asymp n^{-a}, \quad \textit{with } \frac{1}{d+2s} < a < \frac{1}{d+4}. \tag{6.15}$$

Then, for all $x \in \mathbb{T}^d$,

$$\sqrt{nh_n^{d-2}}\big(\widehat{T}_n(x) - T_0(x)\big) \xrightarrow{w} N(0, \Sigma(x)), \quad \textit{as } n \to \infty,$$

where, for all $x \in \mathbb{T}^d$, $\Sigma(x)$ is the positive definite matrix with finite entries given by

$$\Sigma(x) = \frac{1}{p(x)} \int_{\mathbb{R}^d} \xi\xi^\top \left(\frac{\mathcal{F}[K](\mathcal{M}(x)\xi)}{2\pi\langle \mathcal{M}(x)\xi, \xi\rangle}\right)^2 d\xi, \tag{6.16}$$

and $\mathcal{M}(x) = \nabla^2\varphi_0^*(\nabla\varphi_0(x))$.

Theorem 24 shows that the estimator $\widehat{T}_n(x)$ obeys a central limit theorem centered at its population counterpart $T_0(x)$, when the bandwidth h_n lies in the range (6.15). We emphasize again that this range is never empty under the assumption $s > 2$. The lower bound of equation (6.15) implies the undersmoothing condition $h_n = o(h_n^*)$, while the upper bound is needed in order for the error of our linearization of $\widehat{T}_n$ to be of sufficiently low order.

To gain some intuition for the limiting covariance matrix $\Sigma(x)$, it is once again fruitful to compare our result to density estimation. Under the conditions of Theorem 24, it is a simple observation that for all $y \in \mathbb{T}^d$,

$$\sqrt{nh_n^d}\big(\widehat{q}_n(y) - q(y)\big) \xrightarrow{w} N\left(0, q(y)\|K\|_{L^2(\mathbb{R}^d)}^2\right).$$

Thus, the limiting variance of $\widehat{q}_n(y)$ is on the same scale as the $L^2(\mathbb{R}^d)$ norm of the kernel K. In contrast, the limiting covariance of the estimator $\widehat{T}_n(x)$ satisfies

$$\mathrm{tr}(\Sigma(x)) \asymp \frac{1}{p(x)} \int_{\mathbb{R}^d} \left(\frac{\mathcal{F}[K](\xi)}{\|\xi\|}\right)^2 d\xi = \frac{\|K\|_{\dot{H}^{-1}(\mathbb{R}^d)}^2}{p(x)},$$

and is thus on the same scale as the first-order negative Sobolev seminorm of the kernel K. This should not be surprising in view of the formal equivalence

$$\|\widehat{T}_n - T_0\|_{L^2(P)} \asymp \|\widehat{q}_n - q\|_{H^{-1}(\mathbb{T}^d)}$$

which can be anticipated from Theorem 18 of Chapter 5 and remarks thereafter. Furthermore, the following can be said about the off-diagonal entries of $\Sigma(x)$.

Lemma 57. *Assume the same conditions as Theorem 24. Assume further that K is radial, and that $P = Q$. Then, the covariance matrix $\Sigma(x)$ defined in Theorem 24 is diagonal.*

Lemma 57 shows that, for radial kernels, the estimator $\widehat{T}_n(x)$ has asymptotically independent entries when $P = Q$. This property typically fails to hold when $P \neq Q$, however.

6.2.4 Failure of Weak Convergence

We have shown that the sequence $\widehat{T}_n - T_0$ enjoys a pointwise central limit theorem under suitable conditions. We next state a negative result, showing that, under identical conditions, the process $\widehat{T}_n - T_0$ does not converge weakly in $L^2(\mathbb{T}^d)$ to a non-degenerate limit, when $d \geq 3$. More precisely, we will show that the process $\widehat{\varphi}_n - \varphi_0$ does not converge to a non-degenerate limit weakly in $H_0^1(\mathbb{T}^d)$.

Theorem 25. *Assume the same conditions as Theorem 24. Let $(\alpha_n)_{n\geq 1}$ be a positive sequence, and define the process*

$$\mathbb{G}_n = \alpha_n(\widehat{\varphi}_n - \varphi_0), \quad n = 1, 2, \ldots$$

viewed as a random element in $H_0^1(\mathbb{T}^d)$. Then, the following hold.

(i) If $\alpha_n = o(\sqrt{nh_n^{d-2}})$, then $\mathbb{G}_n$ converges weakly to 0 in $H_0^1(\mathbb{T}^d)$.

(ii) If $\alpha_n \gtrsim \sqrt{nh_n^{d-2}}$, then $\mathbb{G}_n$ does not converge weakly in $H_0^1(\mathbb{T}^d)$.

We prove Theorem 25 in Section 6.5.4, by showing that the $H_0^1(\mathbb{T}^d)$ projection of $\mathbb{G}_n$ along any fixed direction typically vanishes at a significantly faster rate than the convergence rate of $\mathbb{G}_n$ under the $H_0^1(\mathbb{T}^d)$ norm. To establish this result, we require the same conditions on the bandwidth h_n as in Theorem 24. We do not rule out the possibility that $\mathbb{G}_n$ could converge weakly when h_n falls outside of this range, and in particular when $h_n = o(n^{-1/(d+s+1)})$.

6.3 Quantitative Linearization of the Monge-Ampère Equation

The aim of this section is to prove Theorem 23 and Proposition 36. We begin by stating several important properties of the operator L.

6.3.1 The Differential Operator *L*

Let $p, q \in \mathcal{C}_+^{2+\beta}(\mathbb{T}^d)$, where the constant β was fixed in equation (6.7), and define the operator

$$L : H_0^2(\mathbb{T}^d) \to L_0^2(\mathbb{T}^d), \quad Lu = -\mathrm{div}(q\nabla u(\nabla\varphi_0^*)).$$

A simple calculation reveals that the $L^2(\mathbb{T}^d)$ adjoint of L is given by

$$L^* : H_0^2(\mathbb{T}^d) \to L_0^2(\mathbb{T}^d), \quad L^* v = -\mathrm{div}(p\nabla v(\nabla\varphi_0)),$$

which implies, in particular, that L is self-adjoint if and only if $P = Q$. It will be convenient to note that L is closely-related to a distinct operator E, which in turn is self-adjoint even when $P \neq Q$. This operator is defined by

$$E : H_0^2(\mathbb{T}^d) \to L_0^2(\mathbb{T}^d), \quad Eu = -\mathrm{div}(A\nabla u),$$

where we fix once and for all the matrix-valued map

$$A(x) = p(x)\nabla^2\varphi_0^*(\nabla\varphi_0(x)), \quad x \in \mathbb{T}^d.$$

The relation between L and E is described next.

Lemma 58. *Let $p, q \in \mathcal{C}_+^{2+\beta}(\mathbb{T}^d)$. Then, for any $u \in \mathcal{C}_0^2(\mathbb{T}^d)$, it holds that*

$$Eu = L[u](\nabla\varphi_0)\det(\nabla^2\varphi_0), \quad \textit{and} \quad Lu = E[u](\nabla\varphi_0^*)\det(\nabla^2\varphi_0^*). \tag{6.17}$$

Furthermore,

$$Eu = -\det(\nabla^2\varphi_0)\Big\{q(\nabla\varphi_0)\langle(\nabla^2\varphi_0)^{-1}, \nabla^2 u\rangle + \langle\nabla q(\nabla\varphi_0), \nabla u\rangle\Big\}. \tag{6.18}$$

Proof of Lemma 58. For any $u \in H_0^2(\mathbb{T}^d)$ and any test function $v \in \mathcal{C}_0^\infty(\mathbb{T}^d)$, it follows by integration by parts that

$$\langle Eu, v\rangle_{L^2(\mathbb{T}^d)} = \int_{\mathbb{T}^d} \nabla v^\top \nabla^2\varphi_0^*(\nabla\varphi_0)\nabla u dP = \int_{\mathbb{T}^d} \nabla v(\nabla\varphi_0^*)^\top \nabla^2\varphi_0^*\nabla u(\nabla\varphi_0^*) dQ,$$

where the final inequality follows by a change of variable. Deduce that

$$\langle Eu, v\rangle_{L^2(\mathbb{T}^d)} = \int_{\mathbb{T}^d} \nabla[v(\nabla\varphi_0^*)]^\top \nabla u(\nabla\varphi_0^*) dQ = \int_{\mathbb{T}^d} v(\nabla\varphi_0^*) Lu,$$

where we again integrated by parts. The above is equivalent to

$$\langle Eu, v\rangle_{L^2(\mathbb{T}^d)} = \langle v, L[u](\nabla\varphi_0)\det(\nabla^2\varphi_0)\rangle_{L^2(\mathbb{T}^d)},$$

which implies the first claim of equation (6.17). The second claim follows analogously. To deduce equation (6.18), note that we may expand L as

$$Lu = -q\langle\nabla^2\varphi_0^*, \nabla^2 u(\nabla\varphi_0^*)\rangle - \langle\nabla q, \nabla u(\nabla\varphi_0^*)\rangle,$$

thus

$$L[u](\nabla\varphi_0) = -q(\nabla\varphi_0)\langle\nabla^2\varphi_0^*(\nabla\varphi_0), \nabla^2 u\rangle - \langle\nabla q(\nabla\varphi_0), \nabla u\rangle.$$

Since $\nabla\varphi_0^*$ is the inverse of $\nabla\varphi_0$, it holds that $(\nabla^2\varphi_0)^{-1} = \nabla^2\varphi_0^*(\nabla\varphi_0)$, and claim (6.18) now follows from claim (6.17). □

With Lemma 58 in place, the properties of the operator L can be deduced from the standard theory of elliptic PDEs subject to periodic boundary conditions, which we summarize in Appendix 6.A. Indeed, under the smoothness assumption $p, q \in \mathcal{C}^{2+\beta}_+(\mathbb{T}^d)$, it follows from Theorem 3 that the eigenvalues of $\nabla^2 \varphi_0^*$, and hence of the matrix A, are uniformly bounded from below over $\mathbb{T}^d$ by positive constants depending on $\omega_{2+\beta}(p,q)$. The operator E is therefore uniformly elliptic. Furthermore, the entries of A lie in $\mathcal{C}^{1+\beta}(\mathbb{T}^d)$, thus it follows from Lemma 77 in Appendix 6.A that the mapping $E : H^2_0(\mathbb{T}^d) \to L^2_0(\mathbb{T}^d)$ is a bijection, whose restriction to $\mathcal{C}^{2+\beta}_0(\mathbb{T}^d)$ is a bijection onto $\mathcal{C}^\beta_0(\mathbb{T}^d)$. Additionally, the following norm equivalences hold:

$$\|Eu\|_{L^2(\mathbb{T}^d)} \asymp \|u\|_{H^2(\mathbb{T}^d)}, \quad \|Eu\|_{\mathcal{C}^{2+\beta}(\mathbb{T}^d)} \asymp \|u\|_{\mathcal{C}^{\beta}(\mathbb{T}^d)}, \tag{6.19}$$

where the implicit constants depend only on $\omega_{2+\beta}(p,q), d, \beta$. From here, we may deduce the following.

Lemma 59. *Let $p, q \in \mathcal{C}^{2+\beta}_+(\mathbb{T}^d)$. Then, the mapping $L : H^2_0(\mathbb{T}^d) \to L^2_0(\mathbb{T}^d)$ is a bijection, whose restriction to $\mathcal{C}^{2+\beta}_0(\mathbb{T}^d)$ is a bijection onto $\mathcal{C}^\beta_0(\mathbb{T}^d)$, and satisfies the norm equivalences*

$$\|Lu\|_{L^2(\mathbb{T}^d)} \asymp \|u\|_{H^2(\mathbb{T}^d)}, \quad \|Lu\|_{\mathcal{C}^{\beta}(\mathbb{T}^d)} \asymp \|u\|_{\mathcal{C}^{2+\beta}(\mathbb{T}^d)}.$$

Furthermore, for all $f \in \mathcal{C}^\beta_0(\mathbb{T}^d)$, it holds

$$E^{-1} f = L^{-1}[f(\nabla \varphi_0^*)\det(\nabla^2 \varphi_0^*)], \quad \textit{and} \quad L^{-1} f = E^{-1}[f(\nabla\varphi_0)\det(\nabla^2\varphi_0)].$$

By reasoning as in Appendix 6.A, it can be seen that for any $f \in L^2_0(\mathbb{T}^d)$, the unique map $u \in H^2_0(\mathbb{T}^d)$ which solves the equation $Lu = f$ over $\mathbb{T}^d$ satisfies the identity

$$\langle u, v\rangle_A = \langle f, v(\nabla\varphi_0^*)\rangle_{L^2(\mathbb{T}^d)}, \quad \text{for all } v \in H^1_0(\mathbb{T}^d), \tag{6.20}$$

where we define the bilinear form

$$\langle u, v\rangle_A := \int_{\mathbb{T}^d} \langle A\nabla u, \nabla v\rangle d\mathcal{L} = \int_{\mathbb{T}^d} \langle \nabla\varphi_0^*(\nabla\varphi_0)\nabla u, \nabla v\rangle dP, \quad \text{for all } u, v \in H^1_0(\mathbb{T}^d).$$

Using again the assumption that $p, q \in \mathcal{C}^{2+\beta}_+(\mathbb{T}^d)$, and hence that A has its eigenvalues bounded from above and below by positive constants over $\mathbb{T}^d$, the above bilinear form defines an inner product on $H^1_0(\mathbb{T}^d)$, which is equivalent to the standard inner product $\langle\cdot,\cdot\rangle_{H^1(\mathbb{T}^d)}$. The characterization (6.20) implies the following simple norm equivalence which will be used repeatedly, and which we prove for completeness in Appendix 6.B.1.

Lemma 60. *Let $p, q \in \mathcal{C}^{2+\beta}_+(\mathbb{T}^d)$. Then, for all $u \in H^2_0(\mathbb{T}^d)$,*

$$\|Lu\|_{H^{-1}(\mathbb{T}^d)} \asymp \|u\|_{H^1(\mathbb{T}^d)},$$

where the implicit constants depend only on $\omega_{2+\beta}(p,q), d, \beta$.

With these properties in place, we turn to the proof of Theorem 23.

6.3.2 Linearization of Monge-Ampère: Proof of Theorem 23

Recall that we define $\mathcal{Q} = [0,1]^d$. Throughout the proof, let $C = C(\omega_{2+\beta}(p,q,\widehat{q}), d, \beta) > 0$ denote a constant whose value is permitted to change from line to line. Under the assumption $p, q, \widehat{q} \in \mathcal{C}_+^{2+\beta}(\mathbb{T}^d)$, it follows from Theorem 3 that,

$$\|\widehat{\varphi}\|_{\mathcal{C}^{2+\beta}(\mathcal{Q})} \vee \|\varphi_0\|_{\mathcal{C}^{2+\beta}(\mathcal{Q})} \leq C \tag{6.21}$$

and

$$\nabla^2\widehat{\varphi} \succeq I_d/C, \quad \nabla^2\varphi_0 \succeq I_d/C, \quad \text{over } \mathcal{Q}. \tag{6.22}$$

With the above regularity estimates, we may define the operators

$$\Psi : \mathcal{C}_0^{2+\beta}(\mathcal{Q}) \to \mathcal{C}_0^{\beta}(\mathcal{Q}), \quad \Psi[\varphi] = p - \det(\nabla^2\varphi)q(\nabla\varphi),$$

and

$$\widehat{\Psi} : \mathcal{C}_0^{2+\beta}(\mathcal{Q}) \to \mathcal{C}_0^{\beta}(\mathcal{Q}), \quad \widehat{\Psi}[\varphi] = p - \det(\nabla^2\varphi)\widehat{q}(\nabla\varphi).$$

As shown in the following Lemma, these operators are Fréchet differentiable, with derivatives that are closely related to the operator E.

Lemma 61. *Let $p, q, \widehat{q} \in \mathcal{C}_+^{2+\beta}(\mathbb{T}^d)$. Then, the maps Ψ and $\widehat{\Psi}$ are Fréchet differentiable at any strongly convex function $\varphi \in \mathcal{C}_0^{2+\beta}(\mathcal{Q})$, with Fréchet derivatives respectively given for all $u \in \mathcal{C}_0^{2+\beta}(\mathcal{Q})$ by*

$$\Psi'_\varphi u = -\det(\nabla^2\varphi)\Big[q(\nabla\varphi)\langle(\nabla^2\varphi)^{-1}, \nabla^2 u\rangle + \langle\nabla u, \nabla q(\nabla\varphi)\rangle\Big],$$

and,

$$\widehat{\Psi}'_\varphi u = -\det(\nabla^2\varphi)\Big[\widehat{q}(\nabla\varphi)\langle(\nabla^2\varphi)^{-1}, \nabla^2 u\rangle + \langle\nabla u, \nabla\widehat{q}(\nabla\varphi)\rangle\Big].$$

Furthermore, there exists a constant $C = C(\omega_{2+\beta}(p,q,\widehat{q}), \beta) > 0$ such that for all $u \in \mathcal{C}_0^{2+\beta}(\mathcal{Q})$,

$$\Big\|\Psi[\varphi+u] - \Psi[u] - \Psi'_\varphi[u]\Big\|_{\mathcal{C}^\beta(\mathcal{Q})} \leq C\|u\|^2_{\mathcal{C}^{2+\beta}(\mathcal{Q})}.$$

The above display also continues to hold with Ψ_φ replaced by $\widehat{\Psi}_\varphi$.

The proof of Lemma 61 is elementary, and appears in Appendix 6.B.2. Notice that $\widehat{\Psi}[\widehat{\varphi}] = \Psi[\varphi_0] = 0$, and thus

$$\widehat{\Psi}[\widehat{\varphi}] - \widehat{\Psi}[\varphi_0] = \Psi[\varphi_0] - \widehat{\Psi}[\varphi_0] = (\widehat{q}(\nabla\varphi_0) - q(\nabla\varphi_0))\det(\nabla^2\varphi_0).$$

Thus, applying Lemma 61 under properties (6.21)–(6.22), we deduce that

$$\Big\|\widehat{\Psi}'_{\varphi_0}[\widehat{\varphi} - \varphi_0] - (\widehat{q}(\nabla\varphi_0) - q(\nabla\varphi_0))\det(\nabla^2\varphi_0)\Big\|_{\mathcal{C}^\beta(\mathcal{Q})} \leq C\|\widehat{\varphi} - \varphi_0\|^2_{\mathcal{C}^{2+\beta}(\mathcal{Q})}. \tag{6.23}$$

The following Lemma will allow us to replace $\widehat{\Psi}_{\varphi_0}$ by Ψ_{φ_0} in the above display

Lemma 62. *Assume the same conditions as Theorem 23. Then, there exist a constant* $C = C(\omega_{2+\beta}(p,q,\widehat{q}),d,\beta) > 0$ *such that for all* $u \in \mathcal{C}_0^{2+\beta}(\mathbb{T}^d)$,

$$\big\|\widehat{\Psi}'_{\varphi_0}[u] - \Psi'_{\varphi_0}[u]\big\|_{\mathcal{C}^\beta(\mathbb{T}^d)} \leq C\Big(\|u\|_{\mathcal{C}^{2+\beta}(\mathbb{T}^d)}\|\widehat{q}-q\|_{\mathcal{C}^\beta(\mathbb{T}^d)} + \|u\|_{\mathcal{C}^{1+\beta}(\mathbb{T}^d)}\|\widehat{q}-q\|_{\mathcal{C}^{1+\beta}(\mathbb{T}^d)}\Big).$$

We defer the proof to Appendix 6.B.4. Implicit in the above assertion is the fact that the map $\widehat{\Psi}'_{\varphi_0}[u] - \widehat{\Psi}'_{\varphi_0}[u]$ is $\mathbb{Z}^d$-periodic, which can be seen by direct inspection using Theorem 3. The latter result also implies that the map $u = \widehat{\varphi} - \varphi_0$ is $\mathbb{Z}^d$-periodic, and thus lies in $\mathcal{C}_0^{2+\beta}(\mathbb{T}^d)$ since $\int_{\mathbb{T}^d}(\widehat{\varphi}-\varphi_0)d\mathcal{L} = \int_Q(\widehat{\varphi}-\varphi_0)d\mathcal{L} = 0$. Returning to equation (6.23) and applying Lemma 62, we deduce that

$$\begin{aligned}\Big\|\Psi'_{\varphi_0}[\widehat{\varphi}-\varphi_0] - (\widehat{q}(\nabla\varphi_0) - q(\nabla\varphi_0))\det(\nabla^2\varphi_0)\Big\|_{\mathcal{C}^\beta(\mathbb{T}^d)}& \\ \leq C\Big(\|\widehat{\varphi}-\varphi_0\|^2_{\mathcal{C}^{2+\beta}(\mathbb{T}^d)} + \|\widehat{q}-q\|_{\mathcal{C}^\beta(\mathbb{T}^d)}\|\widehat{\varphi}-\varphi_0\|_{\mathcal{C}^{2+\beta}(\mathbb{T}^d)}& \\ + \|\widehat{\varphi}-\varphi_0\|_{\mathcal{C}^{1+\beta}(\mathbb{T}^d)}\|\widehat{q}-q\|_{\mathcal{C}^{1+\beta}(\mathbb{T}^d)}\Big),&\end{aligned} \tag{6.24}$$

We bound the right-hand side using the following stability bound, whose proof appears in the following subsection.

Proposition 36. Assume the same conditions as Theorem 23. Then, there exists a constant $c = c(\omega_{2+\beta}(p,q,\widehat{q}),d,\beta) > 0$ such that

$$\|\widehat{\varphi}-\varphi_0\|_{\mathcal{C}^{2+\beta}(\mathbb{T}^d)} \leq c\|\widehat{q}-q\|_{\mathcal{C}^\beta(\mathbb{T}^d)}. \tag{6.25}$$

Applying Proposition 36 to equation (6.24), we arrive at

$$\Big\|\Psi'_{\varphi_0}[\widehat{\varphi}-\varphi_0] - (\widehat{q}(\nabla\varphi_0) - q(\nabla\varphi_0))\det(\nabla^2\varphi_0)\Big\|_{\mathcal{C}^\beta(\mathbb{T}^d)} \leq C\|\widehat{q}-q\|_{\mathcal{C}^\beta(\mathbb{T}^d)}\|\widehat{q}-q\|_{\mathcal{C}^{1+\beta}(\mathbb{T}^d)}.$$

By Lemma 58, the above display is equivalent to

$$\big\|E[\widehat{\varphi}-\varphi_0] - \det(\nabla^2\varphi_0)(\widehat{q}(\nabla\varphi_0) - q(\nabla\varphi_0))\big\|_{\mathcal{C}^\beta(\mathbb{T}^d)} \leq C\|\widehat{q}-q\|_{\mathcal{C}^\beta(\mathbb{T}^d)}\|\widehat{q}-q\|_{\mathcal{C}^{1+\beta}(\mathbb{T}^d)}.$$

Now, recall that the restriction of E to $\mathcal{C}_0^{2+\beta}(\mathbb{T}^d)$ is a bijection onto $\mathcal{C}_0^\beta(\mathbb{T}^d)$, with bounded inverse E^{-1}. Furthermore, the function $\det(\nabla^2\varphi_0)(\widehat{q}(\nabla\varphi_0) - q(\nabla\varphi_0))$ is easily seen to have mean zero, and thus lies in $\mathcal{C}_0^\beta(\mathbb{T}^d)$ by assumption. It follows that

$$\big\|\widehat{\varphi}-\varphi_0 - E^{-1}\left[\det(\nabla^2\varphi_0)(\widehat{q}(\nabla\varphi_0) - q(\nabla\varphi_0))\right]\big\|_{\mathcal{C}^{2+\beta}(\mathbb{T}^d)} \leq C\|\widehat{q}-q\|_{\mathcal{C}^\beta(\mathbb{T}^d)}\|\widehat{q}-q\|_{\mathcal{C}^{1+\beta}(\mathbb{T}^d)}.$$

Recalling Lemma 59, we deduce

$$\big\|(\widehat{\varphi}-\varphi_0) - L^{-1}\left[\widehat{q}-q\right]\big\|_{\mathcal{C}^{2+\beta}(\mathbb{T}^d)} \leq C\|\widehat{q}-q\|_{\mathcal{C}^\beta(\mathbb{T}^d)}\|\widehat{q}-q\|_{\mathcal{C}^{1+\beta}(\mathbb{T}^d)}, \tag{6.26}$$

thus proving the claim. □

6.3.3 Hölder Stability Bound: Proof of Proposition 36

Notice that our assumptions once again imply that properties (6.21)–(6.22) hold. As a result, φ_0 and $\widehat{\varphi}$ solve the following Monge-Ampère equations in the classical sense,

$$\det(\nabla^2 \widehat{\varphi}) = \frac{p}{\widehat{q}(\nabla \widehat{\varphi})}, \quad \det(\nabla^2 \varphi_0) = \frac{p}{q(\nabla \varphi_0)}, \quad \text{over } \mathbb{R}^d.$$

We will reason similarly as in Proposition 9.1 of Caffarelli and Cabré (1995) to argue that $\widehat{\varphi} - \varphi_0$ is in fact the solution to a uniformly elliptic equation over $\mathbb{T}^d$. Indeed, by a first-order Taylor expansion of the determinant function, we have over $\mathbb{R}^d$,

$$\begin{aligned}
&\frac{p}{\widehat{q}(\nabla \widehat{\varphi})} - \frac{p}{q(\nabla \varphi_0)} \\
&\quad = \det(\nabla^2 \widehat{\varphi}) - \det(\nabla^2 \varphi_0) \\
&\quad = \int_0^1 \frac{d}{d\lambda} \det\big(\nabla^2 \varphi_0 + \lambda \nabla^2(\widehat{\varphi} - \varphi_0)\big) d\lambda \\
&\quad = \left\langle \nabla^2(\widehat{\varphi} - \varphi_0), \int_0^1 (\nabla^2 \varphi_0 + \lambda \nabla^2(\widehat{\varphi} - \varphi_0))^{-1} \det\big(\nabla^2 \varphi_0 + \lambda \nabla^2(\widehat{\varphi} - \varphi_0)\big) d\lambda \right\rangle,
\end{aligned}$$

which implies that

$$\langle \mathcal{A}, \nabla^2(\widehat{\varphi} - \varphi_0) \rangle = q(\nabla \varphi_0) - \widehat{q}(\nabla \widehat{\varphi}), \tag{6.27}$$

where $\mathcal{A}$ is the matrix-valued map defined by

$$\mathcal{A}(x) = \frac{q(\nabla \varphi_0(x)) \widehat{q}(\nabla \widehat{\varphi}(x))}{p(x)} \int_0^1 \Big((\nabla^2 \varphi_0 + \lambda \nabla^2(\widehat{\varphi} - \varphi_0))^{-1} \det\big(\nabla^2 \varphi_0 + \lambda \nabla^2(\widehat{\varphi} - \varphi_0)\big) \Big)(x) d\lambda,$$

for all $x \in \mathbb{R}^d$. Notice once again that $\widehat{\varphi} - \varphi_0$ is $\mathbb{Z}^d$-periodic by Theorem 3, and likewise, the entries of $\mathcal{A}$ and the right-hand side of (6.27) are $\mathbb{Z}^d$-periodic. Thus, the equality (6.27) defines a second-order elliptic equation over $\mathbb{T}^d$. From equation (6.22), and the boundedness and positivity of the densities $p, q, \widehat{q}$, we have $\mathcal{A} \succeq I_d / C$ over $\mathbb{T}^d$. Therefore, the operator $\langle \mathcal{A}, \nabla^2(\cdot) \rangle$ is uniformly elliptic. Furthermore, using property (6.21) and the regularity of the densities, the map $\mathcal{A}$ satisfies the conditions of the classical interior Schauder estimates (cf. Lemma 25(i)). We deduce,

$$\|\widehat{\varphi} - \varphi_0\|_{\mathcal{C}^{2+\beta}(\mathbb{T}^d)} \lesssim \|\widehat{\varphi} - \varphi_0\|_{\mathcal{C}^{\beta}(\mathbb{T}^d)} + \|\widehat{q}(\nabla \widehat{\varphi}) - q(\nabla \varphi_0)\|_{\mathcal{C}^{\beta}(\mathbb{T}^d)}.$$

Now,

$$\begin{aligned}
\|\widehat{q}(\nabla \widehat{\varphi}) - q(\nabla \varphi_0)\|_{\mathcal{C}^{\beta}(\mathbb{T}^d)} &\leq \|\widehat{q}(\nabla \widehat{\varphi}) - \widehat{q}(\nabla \varphi_0)\|_{\mathcal{C}^{\beta}(\mathbb{T}^d)} + \|\widehat{q}(\nabla \varphi_0) - q(\nabla \varphi_0)\|_{\mathcal{C}^{\beta}(\mathbb{T}^d)} \\
&\lesssim \|\nabla \widehat{\varphi} - \nabla \varphi_0\|_{\mathcal{C}^{\beta}(\mathbb{T}^d)} + \|\widehat{q} - q\|_{\mathcal{C}^{\beta}(\mathbb{T}^d)},
\end{aligned}$$

where the final inequality follows from Lemmas 95–96, where we use in particular the fact that $\widehat{q} \in \mathcal{C}^1(\mathbb{T}^d)$, and that $\nabla \varphi_0$ is a $\mathcal{C}^{1+\beta}(\mathbb{T}^d)$-diffeomorphism. Thus,

$$\|\widehat{\varphi} - \varphi_0\|_{\mathcal{C}^{2+\beta}(\mathbb{T}^d)} \lesssim \|\widehat{\varphi} - \varphi_0\|_{\mathcal{C}^{1+\beta}(\mathbb{T}^d)} + \|\widehat{q} - q\|_{\mathcal{C}^{\beta}(\mathbb{T}^d)}.$$

By the interpolation inequality in Lemma 94, for any $\epsilon > 0$ there exists $C(\epsilon) > 0$ such that

$$\|\widehat{\varphi} - \varphi_0\|_{\mathcal{C}^{1+\beta}(\mathbb{T}^d)} \leq C(\epsilon)\|\widehat{\varphi} - \varphi_0\|_{L^\infty(\mathbb{T}^d)} + \epsilon\|\widehat{\varphi} - \varphi_0\|_{\mathcal{C}^{2+\beta}(\mathbb{T}^d)}.$$

Upon choosing $\epsilon = \epsilon(\omega_{2+\beta}(p, q, \widehat{q}), d, \beta)$ sufficiently small, we deduce from the previous two displays that

$$\|\widehat{\varphi} - \varphi_0\|_{\mathcal{C}^{2+\beta}(\mathbb{T}^d)} \lesssim \|\widehat{\varphi} - \varphi_0\|_{L^\infty(\mathbb{T}^d)} + \|\widehat{q} - q\|_{\mathcal{C}^\beta(\mathbb{T}^d)}. \tag{6.28}$$

Now, using again the aforementioned regularity properties of the map $\mathcal{A}$, and the uniform ellipticity of equation (6.27), we may apply the De Giorgi-Nash-Moser bound of Lemma 76 to obtain that for any fixed $r > 2 \vee d/2$,

$$\begin{aligned}
\|\widehat{\varphi} - \varphi_0\|_{L^\infty(\mathbb{T}^d)} &\lesssim \|\widehat{\varphi} - \varphi_0\|_{L^2(\mathbb{T}^d)} + \|\widehat{q}(\nabla\widehat{\varphi}) - q(\nabla\varphi_0)\|_{L^r(\mathbb{T}^d)} \\
&\lesssim \|\widehat{\varphi} - \varphi_0\|_{L^2(\mathbb{T}^d)} + \|\widehat{q}(\nabla\widehat{\varphi}) - q(\nabla\widehat{\varphi})\|_{L^r(\mathbb{T}^d)} + \|q(\nabla\widehat{\varphi}) - q(\nabla\varphi_0)\|_{L^r(\mathbb{T}^d)} \\
&\leq \|\widehat{\varphi} - \varphi_0\|_{L^2(\mathbb{T}^d)} + \|\widehat{q} - q\|_{L^\infty(\mathbb{T}^d)} + \|q(\nabla\widehat{\varphi}) - q(\nabla\varphi_0)\|_{L^r(\mathbb{T}^d)}.
\end{aligned}$$

Using an $L^r(\mathbb{T}^d)$ interpolation inequality (Lemma 81), one has that for any $\epsilon > 0$, there exists $C'(\epsilon) > 1$ such that

$$\begin{aligned}
\|q(\nabla\widehat{\varphi}) - q(\nabla\varphi_0)\|_{L^r(\mathbb{T}^d)} &\leq C'(\epsilon)\|q(\nabla\widehat{\varphi}) - q(\nabla\varphi_0)\|_{L^2(\mathbb{T}^d)} + \epsilon\|q(\nabla\widehat{\varphi}) - q(\nabla\varphi_0)\|_{L^\infty(\mathbb{T}^d)} \\
&\lesssim C'(\epsilon)\|\nabla\widehat{\varphi} - \nabla\varphi_0\|_{L^2(\mathbb{T}^d)} + \epsilon\|\nabla\widehat{\varphi} - \nabla\varphi_0\|_{L^\infty(\mathbb{T}^d)},
\end{aligned}$$

where we used the fact that $q \in \mathcal{C}^1(\mathbb{T}^d)$ in the final inequality. We have thus shown

$$\begin{aligned}
\|\widehat{\varphi} - \varphi_0\|_{L^\infty(\mathbb{T}^d)} &\lesssim \|\widehat{q} - q\|_{L^\infty(\mathbb{T}^d)} + C'(\epsilon)\|\nabla\widehat{\varphi} - \nabla\varphi_0\|_{L^2(\mathbb{T}^d)} + \epsilon\|\nabla\widehat{\varphi} - \nabla\varphi_0\|_{L^\infty(\mathbb{T}^d)} \\
&\lesssim \|\widehat{q} - q\|_{\mathcal{C}^\beta(\mathbb{T}^d)} + C'(\epsilon)\|\widehat{\varphi} - \varphi_0\|_{H^1(\mathbb{T}^d)} + \epsilon\|\widehat{\varphi} - \varphi_0\|_{\mathcal{C}^{2+\beta}(\mathbb{T}^d)}.
\end{aligned}$$

Returning to equation (6.28), we thus have

$$\|\widehat{\varphi} - \varphi_0\|_{\mathcal{C}^{2+\beta}(\mathbb{T}^d)} \lesssim \|\widehat{q} - q\|_{\mathcal{C}^\beta(\mathbb{T}^d)} + C'(\epsilon)\|\widehat{\varphi} - \varphi_0\|_{H^1(\mathbb{T}^d)} + \epsilon\|\widehat{\varphi} - \varphi_0\|_{\mathcal{C}^{2+\beta}(\mathbb{T}^d)},$$

where the implicit constants in the symbol "$\lesssim$" do not depend on ϵ. We may therefore choose $\epsilon = \epsilon(\omega_{2+\beta}(p, q, \widehat{q}), d, \beta)$ sufficiently small, but fixed, such that

$$\|\widehat{\varphi} - \varphi_0\|_{\mathcal{C}^{2+\beta}(\mathbb{T}^d)} \lesssim \|\widehat{q} - q\|_{\mathcal{C}^\beta(\mathbb{T}^d)} + \|\widehat{\varphi} - \varphi_0\|_{H^1(\mathbb{T}^d)}.$$

Furthermore, by Lemma 18 of Chapter 5 and remarks thereafter, we have

$$\|\widehat{\varphi} - \varphi_0\|_{H^1(\mathbb{T}^d)} \lesssim \|\widehat{q} - q\|_{H^{-1}(\mathbb{T}^d)} \lesssim \|\widehat{q} - q\|_{\mathcal{C}^\beta(\mathbb{T}^d)}. \tag{6.29}$$

We deduce that

$$\|\widehat{\varphi} - \varphi_0\|_{\mathcal{C}^{2+\beta}(\mathbb{T}^d)} \lesssim \|\widehat{q} - q\|_{\mathcal{C}^\beta(\mathbb{T}^d)}, \tag{6.30}$$

as claimed. □

Remark 5 (Different Source Measures)**.** By a simple extension of the preceding proofs, it is also possible to derive a linearization bound under variations of both the source and target measures. Concretely, it can be shown that for any densities $p, q, \widehat{p}, \widehat{q} \in \mathcal{C}_+^{2+\beta}(\mathbb{T}^d)$, if $\overline{\varphi}$ denotes the unique continuously differentiable convex function, with mean zero over $\mathcal{Q}$, whose gradient pushes forward $\widehat{p}$ onto $\widehat{q}$, then, one has

$$\begin{aligned}\Big\|(\overline{\varphi} - \varphi_0) - E^{-1}[\widehat{p} - p] - L^{-1}[\widehat{q} - q]\Big\|_{\mathcal{C}^{2+\beta}(\mathbb{T}^d)} \\ \lesssim \|\widehat{p} - p\|_{\mathcal{C}^{1+\beta}(\mathbb{T}^d)}\|\widehat{p} - p\|_{\mathcal{C}^{\beta}(\mathbb{T}^d)} + \|\widehat{q} - q\|_{\mathcal{C}^{1+\beta}(\mathbb{T}^d)}\|\widehat{q} - q\|_{\mathcal{C}^{\beta}(\mathbb{T}^d)}.\end{aligned}$$

Such a bound may be used to derive the pointwise limiting distribution of the optimal transport map pushing forward a kernel density estimator onto another, which would yield a two-sample analogue of Theorem 24. We omit this extension in the interest of brevity.

6.4 Bias and Variance Bounds

We now return our attention to the estimator $\widehat{T}_n = \nabla\widehat{\varphi}_n$ defined in Section 6.1. Theorem 23 suggests that for any given $x \in \mathbb{T}^d$, the sequence $(\widehat{T}_n - T_0)(x)$ is well-approximated by

$$\nabla L^{-1}[\widehat{q}_n - q](x) = \nabla L^{-1}[\widehat{q}_n - q_{h_n}](x) + \nabla L^{-1}[q_{h_n} - q](x). \tag{6.31}$$

As discussed in Section 6.1, the stochastic term on the right-hand side of the above display will characterize the fluctuations of $\widehat{T}_n(x)$ around its mean, while the deterministic term above will characterizes the bias of $\widehat{T}_n(x)$. We analyze these two quantities next.

By linearity of L, the stochastic term can be decomposed as

$$\nabla L^{-1}[\widehat{q}_n - q_{h_n}](x) = \frac{1}{n}\sum_{i=1}^n \nabla L^{-1}\big[\overline{K}_{h_n}(Y_i - \cdot) - q_{h_n}\big](x), \tag{6.32}$$

where $\overline{K}_{h_n}$ denotes the $\mathbb{Z}^d$-periodization of K_{h_n}, as defined in Appendix A. The following result describes the limiting covariance of the summands in the above display.

Lemma 63. *Let $p, q \in \mathcal{C}_+^{2+\beta}(\mathbb{T}^d)$ for some $\beta \in (0,1)$, and let the kernel K satisfy condition* **K(γ)** *for some $\gamma > 0$. Let $Y \sim Q$. Then, there exist constants $C, \epsilon > 0$ depending only on $\omega_{2+\beta}(p,q), K, d, \beta$ such that for any $x \in \mathbb{T}^d$,*

$$\Big\|h_n^{d-2}\mathit{Cov}\Big\{\nabla L^{-1}\big[\overline{K}_{h_n}(Y - \cdot) - q_{h_n}\big](x)\Big\} - \Sigma(x)\Big\| \leq Ch_n^{\epsilon},$$

where $\Sigma(x)$ is the positive definite matrix defined in equation (6.16).

Lemma 63 implies that the covariance matrix of any given term in the summation (6.32) has norm diverging at the pointwise rate h_n^{2-d}, which coincides with the corresponding $L^2(\mathbb{T}^d)$ rate of divergence. Indeed, by Lemma 60, it holds that

$$\mathbb{E}\big\|\nabla L^{-1}\big[\overline{K}_{h_n}(Y - \cdot) - q_{h_n}\big]\big\|_{L^2(\mathbb{T}^d)}^2 \asymp \mathbb{E}\big\|\overline{K}_{h_n}(Y - \cdot) - q_{h_n}\big\|_{H^{-1}(\mathbb{T}^d)}^2 \asymp h_n^{2-d},$$

where the final order assessment can be deduced as in the proof of Proposition 43. The proof of Lemma 63 turns out to require a more involved argument, and appears in Section 6.4.1.

Let us now provide a bound on the deterministic term.

Lemma 64. *Let $p, q \in \mathcal{C}^s_+(\mathbb{T}^d)$. Assume that K satisfies condition* **K($s+1$)**. *Then, for all $\epsilon > 0$, there exists $C = C(\omega_s(p,q), K, \epsilon, d, s) > 0$ such that*

$$\left\|\nabla L^{-1}[q_{h_n} - q]\right\|_{L^\infty(\mathbb{T}^d)} \leq C h_n^{s+1-\epsilon}.$$

The proof of Lemma 64 appears in Section 6.4.2, and is a consequence of existing gradient estimates for uniformly elliptic PDEs. Once again, up to the arbitrarily small constant $\epsilon > 0$, the rate $h_n^{s+1-\epsilon}$ in the above display coincides with the corresponding $L^2(\mathbb{T}^d)$ rate of decay: one has, by Lemmas 60 and 99,

$$\left\|\nabla L^{-1}[q_{h_n} - q]\right\|_{L^2(\mathbb{T}^d)} \asymp \|q_{h_n} - q\|_{H^{-1}(\mathbb{T}^d)} \asymp h_n^{s+1},$$

where the final order assessment can be obtained by reasoning as in Proposition 43, under the assumption **K($s+1$)**.

6.4.1 Proof of Lemma 63

Throughout the proof, the implicit constants in the symbols $\lesssim$ and $\asymp$ depend only on $\omega_{2+\beta}(p,q), K, d, \beta$. In particular, such constants do not depend on the variable x. By Theorem 3, there exists $\lambda > 1$ such that $\|\varphi_0\|_{\mathcal{C}^3(\mathbb{T}^d)} \leq \lambda$ and $\nabla^2\varphi_0^* \succeq \lambda^{-1} I_d$, and we fix this value of λ throughout the proof. Fix $x \in \mathbb{T}^d$ throughout what follows, and abbreviate $X_0 = \nabla\varphi_0^*(Y)$. We may assume without loss of generality that the representative of X_0 in $\mathbb{T}^d$ is chosen such that

$$\|X_0 - x\|_{\mathbb{T}^d} = \|X_0 - x\|. \tag{6.33}$$

Next, we abbreviate

$$\Sigma_n(x) = \mathrm{Cov}\big(\nabla L^{-1}[\overline{K}_{h_n}(Y - \cdot) - q_{h_n}](x)\big).$$

Since q_{h_n} is deterministic, and L is linear, we may rewrite $\Sigma_n(x)$ as

$$\Sigma_n(x) = \mathrm{Cov}\big(\nabla L^{-1}[\overline{K}^o_{h_n}(Y - \cdot)](x)\big), \quad \text{with } \overline{K}^o_{h_n} = \overline{K}_{h_n} - 1.$$

Deduce from Lemma 59 that

$$\Sigma_n(x) = \mathrm{Cov}\big(\nabla E^{-1}\left[\det(\nabla^2\varphi_0)\overline{K}^o_{h_n}(Y - \nabla\varphi_0(\cdot))\right](x)\big). \tag{6.34}$$

We will derive the limit of this quantity by approximating the operator E with a constant-coefficient differential operator, and showing that the remainder of this approximation is of low order. We proceed in several steps. The proofs of all intermediary results which follow are relegated to Appendix 6.C.

Step 1: Reduction to a constant-coefficient operator. Using Lemma 58, we may write for any $u \in H_0^2(\mathbb{T}^d)$,

$$Eu = -\langle A, \nabla^2 u\rangle - \langle b, \nabla u\rangle,$$

where

$$A = p\nabla^2\varphi_0^*(\nabla\varphi_0), \quad \text{and} \quad b = \det(\nabla^2\varphi_0)\nabla q(\nabla\varphi_0).$$

Let E_0 be defined as the leading term in E with coefficients "frozen" at the point x, namely:

$$E_0u(y) = -\langle A_0, \nabla^2 u(y)\rangle = -\mathrm{div}(A_0\nabla^2 u(y)), \quad y \in \mathbb{T}^d$$

where $A_0 = A(x)$ is a constant matrix. Furthermore, abbreviate $\mathcal{B} = \nabla^2\varphi_0$ and $\mathcal{B}_0 = \nabla^2\varphi_0(x)$. Define the linear map

$$S_x(y) = \nabla\varphi_0(x) + \nabla^2\varphi_0(x)(y-x), \quad y \in \mathbb{T}^d,$$

and set

$$\begin{aligned} u_n &= E^{-1}\big[\det(\mathcal{B})\overline{K}^o_{h_n}(Y - \nabla\varphi_0(\cdot))\big], \quad U_n = \nabla u_n,\\ v_n &= E_0^{-1}\big[\det(\mathcal{B})\overline{K}^o_{h_n}(Y - \nabla\varphi_0(\cdot))\big], \quad V_n = \nabla v_n,\\ w_n &= E_0^{-1}\big[\det(\mathcal{B}_0)\overline{K}^o_{h_n}(Y - S_x(\cdot))\big], \quad W_n = \nabla w_n, \end{aligned}$$

where we emphasize that each of the functions appearing in square brackets in the above display has mean zero, so that the inverses are well-defined. We have the decomposition

$$\begin{aligned} \Sigma_n(x) &= \mathrm{Cov}[U_n(x)]\\ &= \mathrm{Cov}[W_n(x)] + \mathrm{Cov}[(V_n - W_n)(x)] + \mathrm{Cov}[(U_n - V_n)(x)]\\ &+ 2\mathrm{Cov}[W_n(x), (U_n - W_n)(x)] + 2\mathrm{Cov}[(V_n - W_n)(x), (U_n - V_n)(x)], \end{aligned} \tag{6.35}$$

where for two random vectors A and B, we denote by $\mathrm{Cov}(A,B) = \mathbb{E}[(A - \mathbb{E}[A])(B - \mathbb{E}[B])^\top]$ the cross-covariance between A and B. Fix

$$\mathcal{V}_1 = \mathrm{Cov}[W_n(x)], \quad \mathcal{V}_2 = \mathbb{E}\|(V_n - W_n)(x)\|^2, \quad \mathcal{V}_3 = \mathbb{E}\|(U_n - V_n)(x)\|^2.$$

We will show that for some $\epsilon > 0$,

$$\|h_n^{d-2}\mathcal{V}_1 - \Sigma(x)\| \lesssim h_n^\epsilon, \tag{6.36}$$

and, on the other hand, that $\mathcal{V}_2 \vee \mathcal{V}_3 \lesssim h_n^{2-d+2\beta}$. Using the Cauchy-Schwarz inequality, it will then immediately follow that the three cross-covariances in equation (6.35) are of order at most $h_n^{2-d-\beta}$, and the claim will then follow. We analyze each of the terms $\mathcal{V}_1, \mathcal{V}_2, \mathcal{V}_3$ in turn.

Step 2: Bounding term $\mathcal{V}_1$. Write

$$S_{Y,x}(y) = Y - S_x(y) = Y - \nabla\varphi_0(x) - \mathcal{B}_0(y-x).$$

For this step only, we will require the additional constant-coefficient operator

$$G_0 : H_0^2(\mathbb{T}^d) \to L_0^2(\mathbb{T}^d), \quad G_0 u(y) = -\langle \mathcal{B}_0^2 A_0, \nabla^2 u(y)\rangle = -\langle \mathcal{C}_0, \nabla^2 u(y)\rangle,$$

where $\mathcal{C}_0 = p(x)\nabla^2\varphi_0(x)$. The results of Appendix 6.A imply that the restriction of G_0 to $\mathcal{C}_0^{2+\beta}(\mathbb{T}^d)$ is a bijection onto $\mathcal{C}_0^{\beta}(\mathbb{T}^d)$. This operator is motivated by the following simple observation.

Lemma 65. *For all $f \in \mathcal{C}_0^\beta(\mathbb{T}^d)$ and $y \in \mathbb{T}^d$, it holds that*

$$E_0^{-1}[f(S_{Y,x}(\cdot))](y) = G_0^{-1}[f](S_{Y,x}(y)).$$

The proof appears in Appendix 6.C.1. Deduce from Lemma 65 that for any $f \in \mathcal{C}_0^\beta(\mathbb{T}^d)$ and $y \in \mathbb{T}^d$,

$$\nabla_y E_0^{-1}[f(S_{Y,x}(\cdot))](y) = \nabla_y G_0^{-1}[f](S_{Y,x}(y)) = -\mathcal{B}_0 \nabla G_0^{-1}[f](S_{Y,x}(y)),$$

so that

$$W_n(x) = \nabla E_0^{-1}\big[\det(\mathcal{B}_0)\overline{K}^o_{h_n}(S_{Y,x}(\cdot))\big](x) = -\mathcal{B}_0 \det(\mathcal{B}_0)\nabla G_0^{-1}\big[\overline{K}^o_{h_n}\big](Y - \nabla\varphi_0(x)),$$

whence,

$$\mathcal{V}_1 = \det{}^2(\mathcal{B}_0)\mathcal{B}_0 \mathrm{Cov}\big\{\nabla G_0^{-1}\big[\overline{K}^o_{h_n}\big](Y - \nabla\varphi_0(x))\big\}\mathcal{B}_0^\top. \tag{6.37}$$

The following Lemma shows that the covariance in the above display can be taken with respect to the uniform law, without substantial loss.

Lemma 66. *Let $U \sim \mathcal{L}$ be a uniform random variable on $\mathbb{T}^d$. Then, for some $\epsilon > 0$,*

$$\Big\| \mathcal{V}_1 - q(\nabla\varphi_0(x))\det{}^2(\mathcal{B}_0)\mathcal{B}_0 \mathit{Cov}\big\{\nabla G_0^{-1}\big[\overline{K}^o_{h_n}\big](U)\big\}\mathcal{B}_0^\top \Big\| \lesssim h_n^{2+\epsilon-d}.$$

The proof appears in Appendix 6.C.2. Our aim is now to study the covariance appearing in the above display. Recall that $G_0 = -\langle \mathcal{C}_0, \nabla^2(\cdot)\rangle$. Since $\mathcal{C}_0$ is a constant matrix, a simple derivation shows

$$G_0^{-1}[\overline{K}^o_{h_n}] = \sum_{\xi \in \mathbb{Z}_*^d} \frac{\mathcal{F}[\overline{K}_{h_n}](\xi)}{(2\pi)^2 \langle \mathcal{C}_0 \xi, \xi\rangle} e^{2\pi i \langle \xi, \cdot\rangle}.$$

Since $K \in \mathcal{C}_c^\infty(\mathbb{R}^d)$, the above Fourier series converges absolutely and uniformly for any given n (Stein and Weiss, 1971), and we may differentiate it term-by-term to obtain

$$\nabla G_0^{-1}[\overline{K}^o_{h_n}] = \sum_{\xi \in \mathbb{Z}_*^d} \frac{i\xi \mathcal{F}[\overline{K}_{h_n}](\xi)}{2\pi \langle \mathcal{C}_0 \xi, \xi\rangle} e^{2\pi i \langle \xi, \cdot\rangle} = \sum_{\xi \in \mathbb{Z}_*^d} \frac{i\xi \mathcal{F}[K](h_n \xi)}{2\pi \langle \mathcal{C}_0 \xi, \xi\rangle} e^{2\pi i \langle \xi, \cdot\rangle},$$

where we used the Poisson summation formula in the final equality (cf. equation (B.10)). Again, the above series converges uniformly, and we thus have

$$\mathbb{E}\big[\nabla G_0^{-1}[\overline{K}^o_{h_n}](U)\big] = \sum_{\xi\in\mathbb{Z}^d_*} \frac{i\xi\mathcal{F}[K](h_n\xi)}{2\pi\langle\mathcal{C}_0\xi,\xi\rangle}\int_{\mathbb{T}^d} e^{2\pi i\langle\xi,y\rangle}dy = 0.$$

Deduce that

$$\begin{aligned}
&\mathrm{Cov}\big\{\nabla G_0^{-1}[\overline{K}^o_{h_n}](U)\big\}\\
&\quad= \mathbb{E}\Big\{\big(\nabla G_0^{-1}[\overline{K}^o_{h_n}](U)\big)\overline{\big(\nabla G_0^{-1}[\overline{K}^o_{h_n}](U)\big)}^\top\Big\} \qquad (6.38)\\
&\quad= \mathbb{E}\left\{\left(\sum_{\xi\in\mathbb{Z}^d_*}\frac{i\xi\mathcal{F}[K](h_n\xi)}{2\pi\langle\mathcal{C}_0\xi,\xi\rangle}e^{2\pi i\langle\xi,U\rangle}\right)\left(\sum_{\zeta\in\mathbb{Z}^d_*}\frac{-i\zeta\mathcal{F}[K](h_n\zeta)}{2\pi\langle\mathcal{C}_0\zeta,\zeta\rangle}e^{-2\pi i\langle\zeta,U\rangle}\right)^\top\right\}\\
&\quad= \mathbb{E}\left\{\sum_{\xi,\zeta\in\mathbb{Z}^d_*}\frac{\xi\zeta^\top\mathcal{F}[K](h_n\xi)\mathcal{F}[K](h_n\zeta)}{(2\pi)^2\langle\mathcal{C}_0\xi,\xi\rangle\langle\mathcal{C}_0\zeta,\zeta\rangle}e^{2\pi i\langle\xi-\zeta,U\rangle}\right\}\\
&\quad= \sum_{\xi\in\mathbb{Z}^d_*}\xi\xi^\top\left(\frac{\mathcal{F}[K](h_n\xi)}{2\pi\langle\mathcal{C}_0\xi,\xi\rangle}\right)^2. \qquad (6.39)
\end{aligned}$$

Up to suitable scaling, the expression (6.38) is as a matrix-valued improper Riemann sum, whose limit is described next.

Lemma 67. *There exist constants $c,\epsilon>0$ depending only on d and K such that*

$$\left\|h_n^{d-2}\mathit{Cov}\big\{\nabla G_0^{-1}[\overline{K}^o_{h_n}](U)\big\} - \int_{\mathbb{R}^d}\xi\xi^\top\left(\frac{\mathcal{F}[K](\xi)}{2\pi\langle\mathcal{C}_0\xi,\xi\rangle}\right)^2 d\xi\right\| \leq ch_n^\epsilon.$$

The proof appears in Appendix 6.C.3. By Lemmas 66–67, after possibly decreasing the value of ϵ, we deduce that

$$\big\|h_n^{d-2}\mathcal{V}_1 - \widetilde{\Sigma}(x)\big\| \lesssim h_n^\epsilon,$$

where

$$\widetilde{\Sigma}(x) = q(\nabla\varphi_0(x))\det{}^2(\mathcal{B}_0)\mathcal{B}_0\left(\int_{\mathbb{R}^d}\xi\xi^\top\left(\frac{\mathcal{F}[K](\xi)}{2\pi\langle\mathcal{C}_0\xi,\xi\rangle}\right)^2 d\xi\right)\mathcal{B}_0^\top.$$

To complete our analysis of term $\mathcal{V}_1$, it thus remains to show that $\Sigma(x)=\widetilde{\Sigma}(x)$. Recalling the definition of $\mathcal{C}_0$, we have

$$\widetilde{\Sigma}(x) = \frac{q(\nabla\varphi_0(x))\det^2(\mathcal{B}_0)}{p^2(x)}\int_{\mathbb{R}^d}(\mathcal{B}_0\xi)(\mathcal{B}_0\xi)^\top\left(\frac{\mathcal{F}[K](\xi)}{2\pi\langle\mathcal{B}_0\xi,\xi\rangle}\right)^2 d\xi.$$

Applying the change-of-variable $\zeta=\mathcal{B}_0\xi$, we arrive at

$$\widetilde{\Sigma}(x) = \frac{q(\nabla\varphi_0(x))\det(\mathcal{B}_0)}{p^2(x)}\int_{\mathbb{R}^d}\zeta\zeta^\top\left(\frac{\mathcal{F}[K](\mathcal{B}_0^{-1}\zeta)}{2\pi\langle\zeta,\mathcal{B}_0^{-1}\zeta\rangle}\right)^2 d\zeta$$

$$= \frac{1}{p(x)} \int_{\mathbb{R}^d} \zeta\zeta^\top \left(\frac{\mathcal{F}[K](\mathcal{B}_0^{-1}\zeta)}{2\pi\langle\zeta,\mathcal{B}_0^{-1}\zeta\rangle}\right)^2 d\zeta = \Sigma(x),$$

where we used the Monge-Ampère equation $\det(\nabla^2\varphi_0) = p/q(\nabla\varphi_0)$. We thus conclude that

$$\|h_n^{d-2}\mathcal{V}_1 - \Sigma(x)\| \lesssim h_n^\epsilon.$$

Finally, let us also argue that $\Sigma(x)$ is positive definite. We have, for all $v \in \mathbb{R}^d$,

$$\begin{aligned} v^\top\Sigma(x)v &= \frac{1}{p(x)} \int_{\mathbb{R}^d} (v^\top\zeta)^2 \left(\frac{\mathcal{F}[K](\mathcal{B}_0^{-1}\zeta)}{2\pi\langle\zeta,\mathcal{B}_0^{-1}\zeta\rangle}\right)^2 d\zeta \\ &\geq c\int_{\mathbb{R}^d} (v^\top\zeta)^2\|\zeta\|^{-4}\left(\mathcal{F}[K](\mathcal{B}_0^{-1}\zeta)\right)^2 d\zeta, \end{aligned}$$

for a constant $c > 0$ not depending on v, where we used the fact that $\mathcal{B}_0$ has all eigenvalues lying in the positive interval $[\lambda^{-1}, \lambda]$. By assumption **K(γ)**, there exists $\kappa > 0$ such that

$$|\mathcal{F}[K](\mathcal{B}_0^{-1}\xi) - 1| \leq \kappa\|\mathcal{B}_0^{-1}\xi\|^\gamma \leq \kappa\lambda^\gamma\|\xi\|^\gamma,$$

thus, letting S be the ball $\{\xi : \kappa\lambda^\gamma\|\xi\|^\gamma \leq 1/2\}$, we have

$$v^\top\Sigma(x)v \geq c\int_S (v^\top\zeta)^2\|\zeta\|^{-4}\left(1 - \kappa\|\mathcal{B}_0^{-1}\zeta\|\right)^2 d\zeta \geq \frac{c}{4}\int_S (v^\top\zeta)^2\|\zeta\|^{-4}d\zeta > 0,$$

which implies that $\Sigma(x)$ is positive definite. This concludes Step 2 of the proof.

Step 3: Bounding term $\mathcal{V}_2$. Our aim is now to bound

$$\mathcal{V}_2 = \mathbb{E}\big\|\nabla E_0^{-1}\big[\det(\mathcal{B})\overline{K}^o_{h_n}(Y - \nabla\varphi_0(\cdot)) - \det(\mathcal{B}_0)\overline{K}^o_{h_n}(Y - S_x(\cdot))\big](x)\big\|^2.$$

By Proposition 39, the operator E_0 admits a periodic Green's function $\Gamma_{E_0}(y,x)$, which is continuously differentiable with respect to x away from the diagonal $x = y$, and whose gradient satisfies $\|\nabla_x\Gamma_{E_0}(y,x)\| \lesssim \|x-y\|_{\mathbb{T}^d}^{1-d}$. We may therefore write

$$\mathcal{V}_2 \lesssim \mathcal{V}_{2,1} + \mathcal{V}_{2,2},$$

where

$$\begin{aligned} \mathcal{V}_{2,1} &= \mathbb{E}\left[\left(\int_{\mathbb{T}^d} \|x-y\|_{\mathbb{T}^d}^{1-d}\det(\mathcal{B}_0)\big|\overline{K}^o_{h_n}(Y - \nabla\varphi_0(y)) - \overline{K}^o_{h_n}(Y - S_x(y))\big|dy\right)^2\right], \\ \mathcal{V}_{2,2} &= \mathbb{E}\left[\left(\int_{\mathbb{T}^d} \|x-y\|_{\mathbb{T}^d}^{1-d}\big|\overline{K}^o_{h_n}(Y - \nabla\varphi_0(y))\big| \cdot |\det(\mathcal{B}) - \det(\mathcal{B}_0)|dy\right)^2\right]. \end{aligned}$$

To bound $\mathcal{V}_{2,1}$, we make use of the following observation, whose proof is deferred to Section 6.C.4.

Lemma 68. *It holds that*[2]

$$\operatorname{supp}\left(\overline{K}_{h_n}(Y-\nabla\varphi_0(\cdot))\right) \subseteq B(X_0,\lambda h_n), \quad \textit{and}$$
$$\operatorname{supp}\left(\overline{K}_{h_n}(Y-S_x(\cdot))\right) \subseteq B(S_x^{-1}(Y),\lambda h_n).$$

The proof of Lemma 68 appears in Section 6.C.4. Let

$$B_0 = B(X_0,\lambda h_n) \cup B(S_x^{-1}(Y),\lambda h_n).$$

By Lemma 68, we have

$$\mathcal{V}_{2,1} \lesssim \mathbb{E}\left[\left(\int_{B_0} \|x-y\|_{\mathbb{T}^d}^{1-d}\left|\overline{K}_{h_n}(Y-\nabla\varphi_0(y)) - \overline{K}_{h_n}(Y-S_x(y))\right|dy\right)^2\right].$$

Now, and define the random variable

$$\chi = \begin{cases} 0, & d \leq 6 \text{ and } \|x-X_0\|_{\mathbb{T}^d} \wedge \|x-S_x^{-1}(Y)\|_{\mathbb{T}^d} \geq \alpha_n \\ 1, & \text{otherwise} \end{cases},$$

where, given a fixed scalar $\delta \in (0,1)$, we set $\alpha_n = 2\lambda^3 h_n^{1-\delta}$. We have the decomposition $\mathcal{V}_{2,1} \lesssim \mathcal{V}_{2,1,1} + \mathcal{V}_{2,1,2} + \mathcal{V}_{2,1,3}$, where

$$\begin{aligned}
\mathcal{V}_{2,1,1} &= \mathbb{E}\left[\left(\int_{B_0} \|x-y\|_{\mathbb{T}^d}^{1-d}\left|\overline{K}_{h_n}(Y-\nabla\varphi_0(y)) - \overline{K}_{h_n}(Y-S_x(y))\right|dy\right)^2 \chi\right], \\
\mathcal{V}_{2,1,2} &= \mathbb{E}\left[\left(\int_{B_0} \|x-y\|_{\mathbb{T}^d}^{1-d}\left|\overline{K}_{h_n}(Y-\nabla\varphi_0(y))\right|dy\right)^2 (1-\chi)\right], \\
\mathcal{V}_{2,1,3} &= \mathbb{E}\left[\left(\int_{B_0} \|x-y\|_{\mathbb{T}^d}^{1-d}\left|\overline{K}_{h_n}(Y-S_x(y))\right|dy\right)^2 (1-\chi)\right].
\end{aligned}$$

Let us begin by bounding the first term. Since $\|\overline{K}_{h_n}\|_{\mathcal{C}^1(\mathbb{T}^d)} \lesssim h_n^{-(d+1)}$, we have

$$\begin{aligned}
&\left|\overline{K}_{h_n}(Y-\nabla\varphi_0(y)) - \overline{K}_{h_n}(Y-S_x(y))\right| \\
&\quad \lesssim h_n^{-(d+1)}\|\nabla\varphi_0(y) - \nabla\varphi_0(x) - \nabla^2\varphi_0(x)(y-x)\|_{\mathbb{T}^d} \lesssim h_n^{-(d+1)}\|x-y\|_{\mathbb{T}^d}^2,
\end{aligned}$$

thus we obtain

$$\mathcal{V}_{2,1,1} \lesssim h_n^{-2-2d}\mathbb{E}\left[\left(\int_{B_0} \|x-y\|_{\mathbb{T}^d}^{3-d} dy\right)^2 \chi\right]. \tag{6.40}$$

We now make use of the following Lemma.

[2] By abuse of notation, we say the support of a periodic function is contained in a set $B \subseteq [0,1]^d$ if it is contained in the $\mathbb{Z}^d$-translates of B.

Lemma 69. *Let Z be a random variable taking values in $\mathbb{T}^d$, admitting a density with respect to $\mathcal{L}$ which is bounded from above by some $\gamma > 0$ over $\mathbb{T}^d$. Let $t > 0$. Then, there exists $C = C(\gamma, t, d) > 0$ such that for any $\epsilon \in (0, 1/2)$ and any $y \in \mathbb{T}^d$,*

$$\mathbb{E}\left[\left(\int_{B(Z,\epsilon)} \|y - z\|_{\mathbb{T}^d}^{t-d} dz\right)^2\right] \leq C \begin{cases} \epsilon^{d+2t}, & t < d/2 \\ \epsilon^{2t}, & d/2 \leq t < d \\ \epsilon^{2d}, & d \leq t. \end{cases}$$

The proof appears in Section 6.C.5. Notice that the random variables $\nabla\varphi_0^*(Y)$ and $S_x^{-1}(Y)$ both have probability laws which are absolutely continuous with respect to $\mathcal{L}$, with uniformly upper bounded densities. Returning to equation (6.40), we deduce from Lemma 69 that when $d \geq 7$,

$$\mathcal{V}_{2,1,1} \lesssim h_n^{-2-2d} h_n^{d+6} \asymp h_n^{4-d}.$$

On the other hand, when $d \leq 6$, we have by Lemma 84 that

$$\mathcal{V}_{2,1,1} \lesssim h_n^{4-2d}\mathbb{E}[\chi] \lesssim h^{4-d-\delta d}. \tag{6.41}$$

Let us now bound term $\mathcal{V}_{2,1,2}$, which is only nonzero when $d \leq 6$. By a change of variable, it holds that

$$\mathcal{V}_{2,1,2} \lesssim \mathbb{E}\left[\left(\int_{B(0,\lambda^2 h_n)} \|x - \nabla\varphi_0^*(z - Y)\|_{\mathbb{T}^d}^{1-d} \big|\overline{K}_{h_n}(z)\big| dz\right)^2 (1-\chi)\right].$$

Now, over the event $\chi = 0$, we have for all $z \in B(0, \lambda^2 h_n)$,

$$\begin{aligned}\|x - \nabla\varphi_0^*(z - Y)\|_{\mathbb{T}^d} &\geq \|x - \nabla\varphi_0^*(Y)\|_{\mathbb{T}^d} - \|\nabla\varphi_0^*(Y) - \nabla\varphi_0^*(z - Y)\|_{\mathbb{T}^d} \\ &\geq \|x - \nabla\varphi_0^*(Y)\|_{\mathbb{T}^d} - \lambda^3 h_n \\ &\gtrsim \|x - X_0\|,\end{aligned}$$

where we used equation (6.33). It follows that

$$\begin{aligned}\mathcal{V}_{2,1,2} &\lesssim \mathbb{E}\left[\left(\int_{B(0,\lambda^2 h_n)} \|x - X_0\|^{1-d} \big|\overline{K}_{h_n}(z)\big| dz\right)^2 (1-\chi)\right] \\ &\lesssim \mathbb{E}\left[\left(\|x - X_0\|^{1-d}\right)^2 (1-\chi)\right] \\ &\lesssim \int_{B(x,\alpha_n)^c} \|x - z\|^{2-2d} dz \\ &\lesssim \int_{\alpha_n}^{\sqrt{d}} r^{2-2d} r^{d-1} dr = \int_{\alpha_n}^{\sqrt{d}} r^{1-d} dr \asymp h_n^{(1-\delta)(2-d)}.\end{aligned}$$

A similar bound holds for term $\mathcal{V}_{2,1,3}$. Choosing $\delta = 2\beta$, we have thus shown

$$\mathcal{V}_{2,1} \lesssim h_n^{2-d+2\beta}. \tag{6.42}$$

Next, we bound term $\mathcal{V}_{2,2}$. Reasoning similarly as before, we obtain

$$\begin{aligned}
\mathcal{V}_{2,2} &\lesssim \mathbb{E}\left[\left(\int_{\mathbb{T}^d} \|x-y\|_{\mathbb{T}^d}^{1-d}\big|\overline{K}^o_{h_n}(Y-\nabla\varphi_0(y))\big|\big|\det(\mathcal{B}(y))-\det(\mathcal{B}_0)\big|dy\right)^2\right] \\
&\lesssim \mathbb{E}\left[\left(\int_{B(X_0,\lambda h_n)} \|x-y\|_{\mathbb{T}^d}^{2-d}\big|\overline{K}^o_{h_n}(Y-\nabla\varphi_0(y))\big|dy\right)^2\right]. \\
&\lesssim \mathcal{V}_{2,2,1}+\mathcal{V}_{2,2,2},
\end{aligned}$$

where we again use the decomposition

$$\begin{aligned}
\mathcal{V}_{2,2,1} &= \mathbb{E}\left[\left(\int_{B(X_0,\lambda h_n)} \|x-y\|_{\mathbb{T}^d}^{2-d}\big|\overline{K}^o_{h_n}(Y-\nabla\varphi_0(y))\big|dy\right)^2\chi\right], \\
\mathcal{V}_{2,2,2} &= \mathbb{E}\left[\left(\int_{B(X_0,\lambda h_n)} \|x-y\|_{\mathbb{T}^d}^{2-d}\big|\overline{K}^o_{h_n}(Y-\nabla\varphi_0(y))\big|dy\right)^2(1-\chi)\right].
\end{aligned}$$

We clearly have $\mathcal{V}_{2,2,2}\lesssim\mathcal{V}_{2,1,2}$, thus it remains to bound term $\mathcal{V}_{2,2,1}$. In the regime $d\geq 5$, we may use Lemma 69 to obtain

$$\mathcal{V}_{2,2,1}\lesssim h^{-2d}\mathbb{E}\left[\left(\int_{B(X_0,\lambda h_n)}\|x-y\|_{\mathbb{T}^d}^{2-d}dy\right)^2\right]\lesssim h_n^{4-d}.$$

On the other hand, when $3\leq d\leq 4$, we obtain by Lemma 84 that

$$\begin{aligned}
\mathcal{V}_{2,2,1} &\lesssim h^{-2d}\mathbb{E}\left[\left(\int_{B(X_0,\lambda h_n)}\|x-y\|_{\mathbb{T}^d}^{2-d}dy\right)^2\chi\right] \\
&\lesssim h_n^{-2d+4}\mathbb{E}[\chi]\lesssim h_n^{4-d-\delta d}.
\end{aligned}\tag{6.43}$$

Altogether, we thus obtain

$$\mathcal{V}_2\lesssim h_n^{2-d+2\beta}.\tag{6.44}$$

This concludes Step 3 of the proof.

Step 4: Bounding term $\mathcal{V}_3$. We now wish to show that $\mathcal{V}_3\lesssim h_n^{4-d-2\beta}\vee 1$. Notice first that for all $y\in\mathbb{T}^d$,

$$\begin{aligned}
0 &= E[u_n](y)-E_0[v_n](y) \\
&= -\langle A(y),\nabla^2u_n(y)\rangle+\langle A_0,\nabla^2v_n(y)\rangle-\langle b(y),\nabla u_n(y)\rangle \\
&= \langle A_0-A(y),\nabla^2u_n(y)\rangle-\langle A_0,\nabla^2(u_n-v_n)(y)\rangle-\langle b(y),\nabla u_n(y)\rangle,
\end{aligned}$$

which implies that

$$E_0[u_n - v_n] = \langle A_0 - A, \nabla^2 u_n \rangle + \langle b, \nabla u_n(y) \rangle =: f_n,$$

and hence,

$$(U_n - V_n)(x) = \nabla E_0^{-1}[f_n](x) = \int_{\mathbb{T}^d} \nabla_x \Gamma_{E_0}(y,x) f_n(y) dy.$$

It will thus suffice to bound the expected squared norm of the right-hand side. Using again the fact that $\|\nabla_x \Gamma_{E_0}(y,x)\| \lesssim \|x-y\|_{\mathbb{T}^d}^{1-d}$, we have

$$\begin{aligned} \|(U_n - V_n)(x)\| &= \left\| \int_{\mathbb{T}^d} \nabla_x \Gamma_{E_0}(y,x) f_n(y) dy \right\| \\ &\lesssim \int_{\mathbb{T}^d} \|x-y\|_{\mathbb{T}^d}^{1-d} \big[\|A_0 - A(y)\| \|\nabla^2 u_n(y)\| + \|b(y)\| \|\nabla u_n(y)\| \big] dy \\ &\lesssim \int_{\mathbb{T}^d} \|x-y\|_{\mathbb{T}^d}^{1-d} \big[\|x-y\|_{\mathbb{T}^d} \|\nabla^2 u_n(y)\| + \|\nabla u_n(y)\| \big] dy, \end{aligned}$$

where we used the fact that A has entries in $\mathcal{C}^1(\mathbb{T}^d)$. We decompose the right-hand side of the above display into the terms

$$\mathcal{I}_1 = \int_{\mathbb{T}^d} \|x-y\|_{\mathbb{T}^d}^{2-d} \|\nabla^2 u_n(y)\| dy, \quad \mathcal{I}_2 = \int_{\mathbb{T}^d} \|x-y\|_{\mathbb{T}^d}^{1-d} \|\nabla u_n(y)\| dy.$$

We have

$$\mathcal{V}_3 \lesssim \mathbb{E}[\mathcal{I}_1^2] + \mathbb{E}[\mathcal{I}_2^2],$$

and we now bound the latter two terms in turn. To bound $\mathbb{E}[\mathcal{I}_1^2]$, we will make use of the following result concerning the regularity of u_n.

Lemma 70. *There exists a constant $C = C(\omega_{2+\beta}(p,q), \beta, d) > 0$ such that for all $y \in \mathbb{T}^d$,*

$$\|\nabla^2 u_n(y)\| \leq C_1 \left(h_n^{-(d+\beta)} + \|y - X_0\|_{\mathbb{T}^d}^{-d} \right).$$

The proof appears in Section 6.C.6. From Lemma 70 with $\epsilon = \beta$, we have

$$\begin{aligned} &\mathbb{E}\left[\left(\int_{B(X_0, C_2 h_n)} \|x-y\|_{\mathbb{T}^d}^{2-d} \|\nabla^2 u_n(y)\| dy \right)^2 \right] \\ &\qquad \lesssim h_n^{-2(d+\beta)} \mathbb{E}\left[\left(\int_{B(X_0, C_2 h_n)} \|x-y\|_{\mathbb{T}^d}^{2-d} dy \right)^2 \right] \lesssim h_n^{4-d-2\beta}, \end{aligned}$$

where the final inequality follows from Lemma 69. Write $\mathcal{Q}_z = z + [-1/2, 1/2]^d$ for all $z \in \mathbb{R}^d$. We then have,

$$\mathbb{E}[\mathcal{I}_1^2] \lesssim h_n^{4-d-2\beta} + \mathbb{E}\left[\left(\int_{\mathcal{Q}_{X_0} \setminus B(X_0, C_2 h_n)} \|x-y\|_{\mathbb{T}^d}^{2-d} \|\nabla^2 u_n(y)\| dy \right)^2 \right]. \tag{6.45}$$

Thus, by Lemma 70,

$$\mathbb{E}[\mathcal{I}_1^2] \lesssim h_n^{4-d-2\beta} + \mathbb{E}\left[\left(\int_{\mathcal{Q}_0\setminus B(0,C_2h_n)} \|x-X_0-y\|_{\mathbb{T}^d}^{2-d}\|y\|_{\mathbb{T}^d}^{-d}dy\right)^2\right]. \tag{6.46}$$

Let A_n be the event $\|X_0-x\|_{\mathbb{T}^d} \leq h_n/2$. Notice that for all $y \in \mathbb{T}^d$ such that $\|y\|_{\mathbb{T}^d} > h_n$, it holds

$$\|x-X_0-y\|_{\mathbb{T}^d} \geq \|y\|_{\mathbb{T}^d} - \|x-X_0\|_{\mathbb{T}^d} \geq \|y\|_{\mathbb{T}^d}/2, \tag{6.47}$$

over the event A_n. Therefore, we have,

$$\begin{aligned}\mathbb{E}&\left[\left(\int_{\mathcal{Q}_0\setminus B(C_2h_n)} \|x-X_0-y\|_{\mathbb{T}^d}^{2-d}\|y\|_{\mathbb{T}^d}^{-d}dy\right)^2 I(A_n)\right]\\ &\lesssim \mathbb{E}\left[\left(\int_{\mathcal{Q}_0\setminus B(0,C_2h_n)} \|y\|^{2-2d}dy\right)^2 I(A_n)\right]\\ &\lesssim h_n^d\left(\int_{h_n}^{\sqrt{d}} r^{1-d}dr\right)^2 \lesssim h_n^{4-d}\end{aligned}$$

On the other hand, by Lemma 83,

$$\begin{aligned}\mathbb{E}&\left[\left(\int_{\mathcal{Q}_0\setminus B(0,C_2h_n)} \|x-X_0-y\|_{\mathbb{T}^d}^{2-d}\|y\|_{\mathbb{T}^d}^{-d}dy\right)^2 I(A_n^c)\right]\\ &\lesssim h_n^{-2\beta}\mathbb{E}\left[\left(\int_{\mathbb{T}^d} \|x-X_0-y\|_{\mathbb{T}^d}^{2-d}\|y\|_{\mathbb{T}^d}^{\beta-d}dy\right)^2 I(A_n^c)\right]\\ &\lesssim h_n^{-2\beta}\mathbb{E}\left[\left(\|x-X_0\|^{2+\beta-d}\right)^2 I(A_n^c)\right]\\ &= h_n^{-2\beta}\int_{\mathcal{Q}_x\setminus B(x,h_n/2)} \|x-z\|^{4+2\beta-2d}dQ(z)\\ &= h_n^{-2\beta}\int_{\mathcal{Q}_0\setminus B(0,h_n/2)} \|z\|^{4+2\beta-2d}dz \lesssim h_n^{-2\beta}\int_{h_n/2}^{\sqrt{d}} r^{4+2\beta-2d}r^{d-1}dr \asymp h_n^{4-d-2\beta} \vee 1.\end{aligned}$$

Returning to equation (6.46), we have thus shown

$$\mathbb{E}[\mathcal{I}_1^2] \lesssim h_n^{4-d-2\beta} \vee 1.$$

One can bound $\mathbb{E}[\mathcal{I}_2^2]$ using a similar argument, as shown in the following Lemma.

Lemma 71. *It holds that* $\mathbb{E}[\mathcal{I}_2^2] \lesssim h_n^{4-d-2\beta}$.

The proof of Lemma 71 is deferred to Appendix 6.C.7. Combining these facts, we have shown that $\mathcal{V}_3 \lesssim h_n^{4-d-2\beta}$, and the claim follows. □

6.4.2 Proof of Lemma 64

Let β be chosen as in equation (6.7), and assume $\beta < \epsilon/2$. By Lemma 59,

$$\left\|\nabla L^{-1}[q_{h_n} - q]\right\|_{L^\infty(\mathbb{T}^d)} = \left\|\nabla E^{-1} g_n\right\|_{L^\infty(\mathbb{T}^d)},$$

where $g_n = \det(\nabla^2\varphi_0)(q_{h_n} - q) \circ (\nabla\varphi_0)$ has mean zero over $\mathbb{T}^d$. Let u_n be the unique mean-zero solution to the Poisson equation $-\Delta u_n = g_n$ over $\mathbb{T}^d$. Using the gradient estimate stated in Lemma 25(ii), we have

$$\left\|\nabla E^{-1} g_n\right\|_{L^\infty(\mathbb{T}^d)} \lesssim \|E^{-1}g_n\|_{L^\infty(\mathbb{T}^d)} + \|\nabla u_n\|_{\mathcal{C}^\beta(\mathbb{T}^d)}.$$

By further applying a De Giorgi-Nash-Moser bound (cf. Lemma 76), we deduce

$$\left\|\nabla E^{-1} g_n\right\|_{L^\infty(\mathbb{T}^d)} \lesssim \|E^{-1}g_n\|_{L^2(\mathbb{T}^d)} + \|\nabla u_n\|_{\mathcal{C}^\beta(\mathbb{T}^d)}.$$

Now, using the fact that E^{-1} is self-adjoint, we have

$$\begin{aligned}\|E^{-1}g_n\|^2_{L^2(\mathbb{T}^d)} &= \langle E^{-1}g_n, E^{-1}g_n\rangle_{L^2(\mathbb{T}^d)}\\ &\le \langle E^{-2}g_n, g_n\rangle_{L^2(\mathbb{T}^d)}\\ &\le \|E^{-2}g_n\|_{H^2(\mathbb{T}^d)}\|g_n\|_{H^{-2}(\mathbb{T}^d)} \asymp \|E^{-1}g_n\|_{L^2(\mathbb{T}^d)}\|g_n\|_{H^{-2}(\mathbb{T}^d)},\end{aligned}$$

where the final bound follows from the norm equivalence in equation (6.19). Deduce that $\|E^{-1}g_n\|_{L^2(\mathbb{T}^d)} \lesssim \|g_n\|_{H^{-2}(\mathbb{T}^d)}$. Furthermore, using a Sobolev embedding for periodic spaces (e.g. Corollary 3.5.5 of Schmeisser and Triebel (1987)), we have for any fixed $r > 2d/\beta$,

$$\|\nabla u_n\|_{\mathcal{C}^\beta(\mathbb{T}^d)} \lesssim \|u_n\|_{H^{1+2\beta,r}(\mathbb{T}^d)} = \|g_n\|_{H^{2\beta-1,r}(\mathbb{T}^d)},$$

thus we arrive at

$$\left\|\nabla E^{-1} g_n\right\|_{L^\infty(\mathbb{T}^d)} \lesssim \|g_n\|_{H^{2\beta-1,r}(\mathbb{T}^d)}.$$

To bound the right-hand side, let r' be the Hölder conjugate exponent of r. By Lemma 97,

$$\begin{aligned}\|g_n\|_{H^{2\beta-1,r}(\mathbb{T}^d)} &\asymp \sup_{\substack{v\in H^{1-2\beta,r'}(\mathbb{T}^d)\\ \|v\|_{H^{1-2\beta,r'}(\mathbb{T}^d)}=1}} \int v\det(\nabla^2\varphi_0)(q_{h_n} - q)\circ(\nabla\varphi_0)\\ &= \sup_{\substack{v\in H^{1-2\beta,r'}(\mathbb{T}^d)\\ \|v\|_{H^{1-2\beta,r'}(\mathbb{T}^d)}=1}} \int v(\nabla\varphi_0^*)(q_{h_n} - q)\\ &\lesssim \sup_{\substack{w\in H^{1-2\beta,r'}(\mathbb{T}^d)\\ \|w\|_{H^{1-2\beta,r'}(\mathbb{T}^d)}=1}} \int w\cdot(q_{h_n} - q),\end{aligned}$$

where we used the fact that the map $w = v\circ\nabla\varphi_0^*$ satisfies $\|w\|_{H^{1-2\beta,r'}(\mathbb{T}^d)} \lesssim \|v\|_{H^{1-2\beta,r'}(\mathbb{T}^d)}$ under the regularity conditions we have placed on φ_0, by Lemma 82. It follows that

$$\|g_n\|_{H^{2\beta-1,r}(\mathbb{T}^d)} \lesssim \|q_{h_n} - q\|_{H^{2\beta-1,r}(\mathbb{T}^d)} \lesssim h_n^{s+1-2\beta},$$

where we used Proposition 43 and the assumed properties of K. The claim follows. □

6.5 Proofs of Main Results

In the following subsections, we respectively prove Proposition 35, Theorem 24, Lemma 57, and Theorem 25. The key ingredients are the linearization bound of Theorem 23, the variance bound of Lemma 63, and the bias bounds of Lemma 64.

Let $b \geq 1$ be a constant whose value will be determined below. Let $c = c(\omega_s(p,q), K, b, \beta, s, d) > 0$ be a constant whose value is permitted to vary from line to line. By Lemmas 108–109, and by the assumption $p, q \in \mathcal{C}^s_+(\mathbb{T}^d)$, together with the fact that $s > 2$, it holds for any choice of the free parameter β satisfying equation (6.7) that, with probability at least $1 - (c/n^b)$,

$$\|\widehat{q}_n\|_{\mathcal{C}^{2+\beta}(\mathbb{T}^d)} \leq c, \quad \text{and} \quad \widehat{q}_n \geq 1/c, \ \text{ over } \mathbb{T}^d,$$

and hence, by Theorem 3,

$$\|\widehat{\varphi}_n\|_{\mathcal{C}^{2+\beta}(\mathbb{T}^d)} \leq c, \quad \text{and} \quad \nabla^2\widehat{\varphi}_n \succeq I_d/c, \ \text{ over } \mathbb{T}^d.$$

Over the above high-probability event, $\widehat{T}_n$ is the unique continuous optimal transport map pushing forward P onto $\widehat{Q}_n$. We may therefore apply Theorem 23 to the fitted potential $\widehat{\varphi}_n$ over this event. Thus, with probability at least $1 - (c/n^b)$,

$$\big\|(\widehat{\varphi}_n - \varphi_0) - L^{-1}[\widehat{q}_n - q]\big\|_{\mathcal{C}^{2+\beta}(\mathbb{T}^d)} \leq c\|\widehat{q}_n - q\|_{\mathcal{C}^{\beta}(\mathbb{T}^d)}\|\widehat{q}_n - q\|_{\mathcal{C}^{1+\beta}(\mathbb{T}^d)}.$$

Proposition 42 implies that with probability at least $1 - (c/n^b)$,

$$\|\widehat{q}_n - q\|_{\mathcal{C}^{1+\beta}(\mathbb{T}^d)} \lesssim h_n^{s-1-\beta} + \sqrt{\frac{\log n}{nh_n^{d+2+2\beta}}} \lesssim h_n^{s-1-\beta} + \sqrt{\frac{1}{nh_n^{d+2+4\beta}}},$$

and

$$\|\widehat{q}_n - q\|_{\mathcal{C}^{\beta}(\mathbb{T}^d)} \lesssim h_n^{s-\beta} + \sqrt{\frac{1}{nh_n^{d+4\beta}}}.$$

Combining these facts, we deduce that there exists an event A_n with probability content at least $1 - (c/n^b)$ such that, over A_n,

$$\begin{aligned}&\big\|(\widehat{\varphi}_n - \varphi_0) - L^{-1}\big[\widehat{q}_n - q\big]\big\|_{\mathcal{C}^{2+\beta}(\mathbb{T}^d)} \\ &\lesssim \left(h_n^{s-2\beta} + \sqrt{\frac{1}{nh_n^{d+4\beta}}}\right)\left(h_n^{s-1-\beta} + \sqrt{\frac{1}{nh_n^{d+2+4\beta}}}\right) \asymp h_n^{2s-1-3\beta} + \frac{1}{nh_n^{d+1+4\beta}} =: \delta_n.\end{aligned}$$

Since the operator L^{-1} is linear, we may write

$$L^{-1}\left[\widehat{q}_n - q_{h_n}\right](x) = \frac{1}{n}\sum_{i=1}^n Z_{n,i}(x), \quad \text{where} \quad Z_{n,i}(x) := L^{-1}\Big[\overline{K}_{h_n}(X_i - \cdot) - q_{h_n}\Big](x).$$

We thus have, over the event A_n,

$$\left\| L^{-1}[q_{h_n} - q] + \frac{1}{n}\sum_{i=1}^n Z_{n,i} - (\widehat{\varphi}_n - \varphi_0) \right\|_{\mathcal{C}^{2+\beta}(\mathbb{T}^d)} \lesssim \delta_n, \tag{6.48}$$

and hence also

$$\sup_{x\in\mathbb{T}^d} \left\| \nabla L^{-1}[q_{h_n} - q](x) + \frac{1}{n}\sum_{i=1}^n \nabla Z_{n,i}(x) - (\widehat{T}_n - T_0)(x) \right\| \lesssim \delta_n. \tag{6.49}$$

With this bound in place, we now turn to the proofs of the various results, in turn.

6.5.1 Proof of Proposition 35

Let $x \in \mathbb{T}^d$. To bound the bias of $\widehat{T}_n(x)$, use the decomposition

$$\big\|\mathbb{E}[\widehat{T}_n(x) - T_0(x)]\big\| \leq \big\|\mathbb{E}[(\widehat{T}_n(x) - T_0(x))I_{A_n^c}]\big\| + \big\|\mathbb{E}[(\widehat{T}_n(x) - T_0(x))I_{A_n}]\big\|.$$

For the first term, recall that the definition of $\widehat{T}_n$ implies that $\|\widehat{T}_n(x) - T_0(x)\|$ is uniformly bounded by a constant independent of x, and thus

$$\big\|\mathbb{E}[(\widehat{T}_n(x) - T_0(x))I_{A_n^c}]\big\| \lesssim \mathbb{P}(A_n^{\mathsf{c}}) \lesssim n^{-b}.$$

For the second term, it follows from equation (6.49) that

$$\big\|\mathbb{E}[(\widehat{T}_n(x) - T_0(x))I_{A_n}]\big\| \leq \big\|\mathbb{E}\left[\left(\nabla L^{-1}[q_{h_n} - q](x) + (1/n)\textstyle\sum_{i=1}^n \nabla Z_{n,i}(x)\right) I_{A_n}\right]\big\| + \delta_n.$$

Now, we make use of the following.

Lemma 72. *It holds that*

$$\sup_{x\in\mathbb{T}^d} \|\mathbb{E}[((1/n)\textstyle\sum_{i=1}^n \nabla Z_{n,i}(x))\, I_{A_n}]\| \lesssim n^{a(d+\beta)-b}.$$

The proof of Lemma 72 appears in Appendix 6.D.1. Altogether, we arrive at

$$\begin{aligned}\big\|\mathbb{E}[\widehat{T}_n(x) - T_0(x)]\big\| &\leq \big\|\nabla L^{-1}[q_{h_n} - q](x)\big\| + \delta_n + n^{a(d+\beta)-b} \\ &\lesssim_\epsilon h_n^{s+1-\epsilon} + \frac{1}{nh_n^{d+1+4\beta}} + h_n^{2s-1-3\beta} + n^{a(d+\beta)-b},\end{aligned}$$

using Lemma 64. Since $h_n \asymp n^{-a}$, we may choose b large enough to make the final term of low order. Furthermore, the second term is of lower order than the first if β is chosen sufficiently small in terms of ϵ, due to the condition that $a < (d+s+2)^{-1}$. Likewise, after possibly decreasing β further, the third term is of lower order than the first due to the condition $s > 2$. We thus have

$$\big\|\mathbb{E}[\widehat{T}_n(x) - T_0(x)]\big\| \lesssim_\epsilon h_n^{s+1-\epsilon},$$

thus proving the desired bias bound. Reasoning similarly, we have the following bound on the fluctuations,

$$\begin{aligned}
&\big\|\mathrm{Cov}\big[\widehat{T}_n(x)\big]\big\| \\
&\lesssim \left\|\mathrm{Cov}\left[\frac{1}{n}\sum_{i=1}^n \nabla Z_{n,i}(x)\right]\right\| + \left\|\mathrm{Cov}\left[\widehat{T}_n(x) - \frac{1}{n}\sum_{i=1}^n \nabla Z_{n,i}(x)\right]\right\| \\
&= \frac{1}{n}\left\|\mathrm{Cov}\left[\nabla Z_{n,1}(x)\right]\right\| + \left\|\mathrm{Cov}\left[\widehat{T}_n(x) - T_0(x) - \frac{1}{n}\sum_{i=1}^n \nabla Z_{n,i}(x) - \nabla L^{-1}[q_{h_n} - q](x)\right]\right\| \\
&\lesssim \frac{1}{nh_n^{d-2}} + \delta_n^2 + n^{-b} \\
&= \frac{1}{nh_n^{d-2}} + \left(\frac{1}{nh_n^{d+1+4\beta}} + h_n^{2s-1-3\beta}\right)^2 + n^{-b}
\end{aligned}$$

where we used Lemma 63 and equation (6.49). Under the conditions $(d+4(s-1))^{-1} < a < (d+s+2)^{-1}$, the first term dominates, and the claim follows. □

6.5.2 Proof of Theorem 24

From equation (6.49) and Lemma 64, we obtain that for any fixed $x \in \mathbb{T}^d$ and $\epsilon > 0$,

$$\widehat{T}_n(x) - T_0(x) = \frac{1}{n}\sum_{i=1}^n \nabla Z_{n,i}(x) + O_p(h_n^{s+1-\epsilon} + \delta_n),$$

where the symbols O_p and o_p are to be understood coordinate-wise. Let $\Sigma_n(x) = \mathrm{Cov}[\nabla Z_{n,1}(x)]/n$. We have $\|\Sigma_n(x)\| \asymp h_n^{2-d}/n$ by Lemma 63, thus

$$\begin{aligned}
&\Sigma_n^{-1/2}(x)\big(\widehat{T}_n(x) - T_0(x)\big) \\
&\qquad = \frac{\Sigma_n^{-1/2}(x)}{n}\sum_{i=1}^n \nabla Z_{n,i}(x) + O_p\left(\sqrt{nh_n^{d-2}}\left(h_n^{s+1-\epsilon} + h_n^{2s-1-3\beta} + \frac{1}{nh_n^{d+1+4\beta}}\right)\right).
\end{aligned}$$

Choosing ϵ and β sufficiently small in terms of a, we deduce from condition (6.15) that

$$\Sigma_n^{-1/2}(x)\big(\widehat{T}_n(x) - T_0(x)\big) = \frac{\Sigma_n^{-1/2}(x)}{n}\sum_{i=1}^n \nabla Z_{n,i}(x) + o_p(1). \tag{6.50}$$

We will show that the sample average appearing on the right-hand side of the above display is asymptotically Gaussian, by appealing to Lyapunov's central limit theorem (Lemma 85). We will make use of the following Lemma.

Lemma 73. *For any $\delta > 0$, it holds that*

$$\mathbb{E}\big\|\nabla Z_{n,1}(x)\big\|^{2+\delta} \lesssim h_n^{-\delta(d+\beta)+2-d}.$$

The proof of Lemma 73 is simple, and is deferred to Appendix 6.D.2. From Lemmas 63 and 73, we have

$$\frac{\sum_{i=1}^n \mathbb{E}\|\nabla Z_{n,i}(x)\|^{2+\delta}}{\left(\lambda_{\min}[\sum_{i=1}^n \nabla Z_{n,i}(x)]\right)^{\frac{2+\delta}{2}}} \lesssim \frac{n h_n^{-\delta(d+\beta)+2-d}}{n^{\frac{\delta}{2}+1} h_n^{-(d-2)(2+\delta)/2}(1+o(1))} \lesssim \left(n h_n^{d+2+2\beta}\right)^{-\frac{\delta}{2}}.$$

Since β can be taken to be an arbitrarily small constant, the right-hand side of the above display vanishes under condition (6.15) on h_n. Lyapunov's condition is thus satisfied, so we deduce from Lemma 85 that

$$\frac{\Sigma_n^{-1/2}(x)}{n} \sum_{i=1}^n \nabla Z_{n,i}(x) \xrightarrow{w} N(0, I_d).$$

By Lemma 63, we have $\|n h_n^{d-2} \Sigma_n(x) - \Sigma(x)\| \to 0$, and since $\Sigma(x)$ is positive definite, it follows that $\|(n h_n^{d-2}\Sigma_n(x))^{-1/2} - \Sigma^{-1/2}(x)\| \to 0$. The claim follows from here. □

6.5.3 Proof of Lemma 57

Since K is radial, $\mathcal{F}[K]$ is also radial, thus Σ takes the form

$$\Sigma_0 = \int_{\mathbb{R}^d} g(\xi)\xi\xi^\top d\xi,$$

for a radial Schwartz function g. Consider the entry $(1,2)$ of this matrix:

$$\sigma_{1,2} = \int_{\mathbb{R}^d} g(\xi)\xi_1\xi_2 d\xi.$$

By passing to the spherical coordinates:

$$\begin{aligned}
\xi_1 &= r\cos\theta_1\\
\xi_2 &= r\sin\theta_1\cos\theta_2\\
\xi_3 &= r\sin\theta_1\sin\theta_2\cos\theta_3\\
&\vdots\\
\xi_{d-1} &= r\sin\theta_1\ldots\sin\theta_{d-2}\cos\theta_{d-1}\\
\xi_d &= r\sin\theta_1\ldots\sin\theta_{d-2}\sin\theta_{d-1},
\end{aligned}$$

with $\theta_1, \ldots, \theta_{d-2} \in [0,\pi]$, $\theta_{d-1} \in [0, 2\pi]$, $r \geq 0$, and using Jacobian identity

$$dx = (r^{d-1}\sin^{d-2}\theta_1 \sin^{d-3}\theta_2 \ldots \sin\theta_{d-2}) d\theta_n \ldots d\theta_1 dr,$$

we obtain:

$$\begin{aligned}
\sigma_{1,2}^2 = \int_0^\infty \int_0^\pi \cdots \int_0^\pi \int_0^{2\pi} g(r)(r\cos\theta_1)\times&\\
\times (r\sin\theta_1\cos\theta_2)(r^{d-1}\sin^{d-2}\theta_1 \sin^{d-3}\theta_2 \ldots \sin\theta_{d-2}) d\theta_{d-1}\ldots d\theta_1 dr.&
\end{aligned}$$

Notice that the integral with respect to θ_2 in the above display vanishes, i.e.

$$\int_0^\pi \cos\theta_2 \sin^{d-3}\theta_2 d\theta_2 = 0,$$

thus we obtain $\sigma_{1,2}^2 = 0$. A similar argument shows that the other off-diagonal entries of Σ_0 vanish, and the claim follows. □

6.5.4 Proof of Theorem 25

The proof is inspired by Nishiyama (2011) and Stupfler (2014). We makes use of the following two propositions, which may be of independent interest. The first provides upper bounds on projections of $\widehat{\varphi}_n - \varphi_0$ under the inner product $\langle\cdot,\cdot\rangle_A$ defined in Section 6.3.

Proposition 37. Suppose $p, q \in \mathcal{C}_+^s(\mathbb{T}^d)$. Assume that K satisfies condition **K($s+1$)**. Then, for any given $\epsilon > 0$ and $v \in H_0^1(\mathbb{T}^d)$,

$$\langle\widehat{\varphi}_n - \varphi_0, v\rangle_A = O_p\left(n^{-1/2} + h_n^{s+1} + \frac{1}{nh_n^{d+\epsilon}}\right).$$

The proof appears in Appendix 6.D.3. Next, we state a bound on the $L^2(\mathbb{T}^d)$ convergence rate of $\widehat{T}_n$, which is essentially already contained in Chapter 5. Let $\overline{T}_{h_n}$ be the unique continuous optimal transport map pushing P forward onto the probability law Q_{h_n} with density $q_{h_n} = \mathbb{E}[\widehat{q}_n(\cdot)]$.

Proposition 38. Under the same conditions as Proposition 35, it holds that

$$\mathbb{E}\|\widehat{T}_n - \overline{T}_{h_n}\|_{L^2(\mathbb{T}^d)}^2 \asymp \frac{h_n^{2-d}}{n}, \quad \text{and} \quad \left\|\overline{T}_{h_n} - T_0\right\|_{L^2(\mathbb{T}^d)} \lesssim h_n^{s+1}, \tag{6.51}$$

and for any $r > 2$,

$$\mathbb{E}\|\widehat{T}_n - \overline{T}_{h_n}\|_{L^2(\mathbb{T}^d)}^r \lesssim n^{-\frac{r}{2}} h_n^{r\left(1-\frac{d}{2}\right)}.$$

With these two propositions in place, we turn to the proof. We will begin by proving the following.

Lemma 74. *Suppose there exists $\epsilon > 0$ such that*

$$\alpha_n\left(\frac{1}{\sqrt{n}} + h_n^{s+1} + \frac{1}{nh_n^{d+1+\epsilon}}\right) = o(1),$$

and that $\mathbb{G}_n$ converges weakly in $H_0^1(\mathbb{T}^d)$ to a tight random element $\mathbb{G}$ in $H_0^1(\mathbb{T}^d)$. Then, $\mathbb{G} = 0$.

To prove Lemma 74, let us recall some standard facts about the spectral properties of E. Using the norm equivalence (6.19), and the fact that $H_0^1(\mathbb{T}^d)$ is compactly embedded in $L_0^2(\mathbb{T}^d)$ by the Rellich–Kondrachov theorem, it can be seen that E^{-1} is a compact operator when viewed as a map from $L_0^2(\mathbb{T}^d)$ into itself. Furthermore, we have already noted that E^{-1} is self-adjoint and positive definite, so the spectral theorem for compact and self-adjoint operators implies that E^{-1} admits a discrete spectrum, which in turn also implies that E has a discrete spectrum

$$0 < \lambda_1 \leq \lambda_2 \leq \dots$$

with a corresponding sequence of eigenfunctions $\bar{\eta}_1, \bar{\eta}_2, \dots \in L_0^2(\mathbb{T}^d)$ satisfying $E\bar{\eta}_j = \lambda_j \bar{\eta}_j$ for all $j \geq 1$, and forming an orthonormal basis of $L_0^2(\mathbb{T}^d)$. Furthermore, it follows from Lemma 25(i) that $\bar{\eta}_j \in \mathcal{C}_0^2(\mathbb{T}^d)$ for all $j \geq 1$.

Now, define

$$\eta_j = \bar{\eta}_j / \|\bar{\eta}_j\|_A, \quad j \geq 1.$$

We claim that $\{\eta_j\}_{j\geq 1}$ forms an orthonormal basis of $H_0^1(\mathbb{T}^d)$, when the latter is endowed with the inner product $\langle \cdot, \cdot \rangle_A$. This system is clearly dense in $H_0^1(\mathbb{T}^d)$, since $\{\bar{\eta}_j\}$ is dense in $L_0^2(\mathbb{T}^d)$. It is also clearly normalized. Furthermore, note that for all $j \neq k$,

$$\langle \eta_j, \eta_k \rangle_A = \langle E\eta_j, \eta_k \rangle_{L^2(\mathbb{T}^d)} = \lambda_j \left\langle \eta_j, \eta_k \right\rangle_{L^2(\mathbb{T}^d)} = 0.$$

Thus, $\{\eta_j\}$ is indeed an orthonormal basis of $H_0^1(\mathbb{T}^d)$.

Since the elements of this basis belong to $H_0^1(\mathbb{T}^d)$, Proposition 37 implies that for any fixed $j \geq 1$, we have

$$\langle \mathbb{G}_n, \eta_j \rangle_A = O_p\left(\alpha_n \left(\frac{1}{\sqrt{n}} + h_n^{s+1} + \frac{1}{nh_n^{d+1+\epsilon}}\right)\right) = o_p(1),$$

where the implicit constants depend, in particular, on ϵ, j. On the other hand, by the weak convergence of $\mathbb{G}_n$ to $\mathbb{G}$ in the Hilbert space $H_0^1(\mathbb{T}^d)$, endowed with the inner product $\langle \cdot, \cdot \rangle_A$, we have

$$\langle \mathbb{G}_n, \eta_j \rangle_A \xrightarrow{w} \langle \mathbb{G}, \eta_j \rangle_A, \quad j = 1, 2, \dots$$

(see van der Vaart and Wellner (1996), Theorem 1.8.4.). The preceding two displays imply that $\langle \mathbb{G}, \eta_j \rangle_A = 0$, for any $j \geq 1$. Since $\{\eta_j\}_{j\geq 1}$ forms a basis of $H_0^1(\mathbb{T}^d)$, it readily follows that $\mathbb{G} = 0$, thus proving the Lemma.

Let us now show why the Lemma implies the claim. On the one hand, if $\alpha_n = o(\sqrt{nh_n^{d-2}})$, then we have by Proposition 38,

$$\mathbb{E}\|\mathbb{G}_n\|_{H^1(\mathbb{T}^d)} \asymp \alpha_n \mathbb{E}\|\widehat{T}_n - T_0\|_{L^2(\mathbb{T}^d)} = o\left(\left(\sqrt{nh_n^{d-2}}\right)\left(\frac{1}{\sqrt{nh_n^{d-2}}} + h_n^{s+1}\right)\right).$$

Under condition (6.15) on h_n, we have $h_n = o(h_n^*)$ with $h_n^* = n^{-1/(d+2s)}$. We deduce that $\mathbb{E}\|\mathbb{G}_n\|_{H^1(\mathbb{T}^d)} = o(1)$, thus $\mathbb{G}_n$ must converge weakly to 0 in $H_0^1(\mathbb{T}^d)$.

On the other hand, if $\alpha_n \asymp \sqrt{nh_n^{d-2}}$, then we have by Proposition 38,

$$\mathbb{E}\|\mathbb{G}_n\|_{H^1(\mathbb{T}^d)}^2 \gtrsim \frac{\alpha_n^2}{nh_n^{d-2}} \gtrsim 1, \tag{6.52}$$

where we used again the fact that $h_n = o(h_n^*)$. Now, suppose by way of a contradiction that $\mathbb{G}_n$ converges weakly to a tight random element $\mathbb{G}$ taking values in $H_0^1(\mathbb{T}^d)$. Under condition (6.15) on the bandwidth h_n, there exists $\epsilon > 0$ such that the sequence $\alpha_n \asymp \sqrt{nh_n^{d-2}}$ fulfills the assumption of Lemma 74, thus $\mathbb{G}$ must be zero. Now, for this choice of sequence α_n, it follows from Proposition 38 that $\mathbb{E}\|\mathbb{G}_n\|_{H^1(\mathbb{T}^d)}^3$ is uniformly bounded, and thus that the collection $\{\|\mathbb{G}_n\|_{H^1(\mathbb{T}^d)}^2 : n \geq 1\}$ is uniformly integrable. Since $\mathbb{G}_n$ converges weakly to zero, we must then have (van der Vaart and Wellner, 1996) that $\mathbb{E}\|\mathbb{G}_n\|_{H^1(\mathbb{T}^d)}^2 \to 0$, which contradicts equation (6.52). This proves that for $\alpha_n \asymp \sqrt{nh_n^{d-2}}$, the process $\mathbb{G}_n$ does not converge weakly.

Finally, suppose $\alpha_n/\sqrt{nh_n^{d-2}} \to \infty$. If $\mathbb{G}_n$ were to converge weakly in $H_0^1(\mathbb{T}^d)$, then it is clear that the process $\sqrt{nh_n^{d-2}}\mathbb{G}_n$ would have to converge weakly to zero, which contradicts what we have already shown. The claim follows. □

6.A Background on Elliptic PDE

We next summarize several results from the classical theory of uniformly elliptic partial differential equations, which are used throughout our development. Given maps $\mathcal{A} : \mathbb{R}^d \to \mathbb{R}^{d\times d}$ and $b : \mathbb{R}^d \to \mathbb{R}^d$, a second-order differential operator of the form

$$Mu = -\langle \mathcal{A}, \nabla^2 u\rangle - \langle b, \nabla u\rangle$$

is said to be uniformly elliptic over a domain $\Omega \subseteq \mathbb{R}^d$ if it holds that

$$\mathcal{A}(x) \succeq I_d/\lambda, \quad \text{over } \Omega, \tag{6.53}$$

for some $\lambda > 0$. We will primarily be interested in operators M for which the coefficients $\mathcal{A}$ and b have periodic entries, and for which u is subject to periodic boundary conditions. Nevertheless, we begin by stating in full generality several a priori regularity estimates for solutions to equations of the form $Mu = f$. Some of the assumptions in the following statement are stronger than necessary, but sufficient for our purposes.

Lemma 75. *Let Ω be an open set with $\mathrm{diam}(\Omega) \leq D < \infty$, and let $\Omega_0 \subset\subset \Omega$ be an open set. Let $\eta = \mathrm{dist}(\Omega_0, \partial\Omega)$. Let $f \in \mathcal{C}^\beta(\overline{\Omega})$, and suppose $u \in \mathcal{C}^{2+\beta}(\Omega)$ is a solution to the equation*

$$Mu = f \quad \textit{over}\, \Omega,$$

where we assume that the coordinates of $\mathcal{A}$ and b satisfy

$$\max_{1\leq i,j\leq d} \|\mathcal{A}_{ij}\|_{\mathcal{C}^\beta(\overline{\Omega})} \vee \|b_i\|_{\mathcal{C}^\beta(\overline{\Omega})} \leq \lambda \tag{6.54}$$

and that equation (6.53) *holds, for some $\lambda > 0$. Then, the following assertions hold.*

(i) (Schauder Estimate) There exists a constant $C = C(D, d, \lambda, \beta)$ such that

$$\eta^2\|u\|_{\mathcal{C}^2(\Omega_0)} + \eta^{2+\beta}\|u\|_{\mathcal{C}^{2+\beta}(\Omega_0)} \leq C\Big(\|u\|_{L^\infty(\Omega)} + \|f\|_{\mathcal{C}^\beta(\Omega)}\Big).$$

(ii) *(Gradient Estimate) Suppose further that f takes the form $f = g + \mathrm{div}(G)$, where $g \in \mathcal{C}^\beta(\Omega)$ and $G : \Omega \to \mathbb{R}^d$ is a vector field with entries in $\mathcal{C}^{1+\beta}(\Omega)$. Then, there exists a constant $C = C(D, d, \lambda, \beta, \eta) > 0$ such that*

$$\|u\|_{\mathcal{C}^{1+\beta}(\Omega_0)} \leq C\Big(\|u\|_{L^\infty(\Omega)} + \|g\|_{L^\infty(\Omega)} + \|G\|_{\mathcal{C}^\beta(\Omega)}\Big).$$

The bounds of Lemma 75 are standard, and can be deduced from Corollary 6.3 and Theorem 8.32 of Gilbarg and Trudinger (2001). We further make use of the following De Giorgi-Nash-Moser estimate, as stated in Theorem 8.17 of Gilbarg and Trudinger (2001).

Lemma 76. *Let Ω be an open set with $\mathrm{diam}(\Omega) \leq D < \infty$. Assume the coefficients $\mathcal{A}$ and b satisfy conditions (6.53)–(6.54). Given $r > d/2$, let $g \in L^r(\Omega)$ and let $G : \Omega \to \mathbb{R}^d$ be a vector field with entries in $L^{2r}(\Omega)$. Suppose $u \in H_0^1(\Omega)$ is a weak solution to the equation*

$$Mu = g + \mathrm{div}(G), \quad \text{in } \Omega.$$

Then, there exists $C = C(D, M, r, \rho, \lambda, \beta) > 0$ such that for all $R > 0$, all balls $B_{2R} := B(y, 2R) \subseteq \Omega$ and all $\rho > 1$,

$$\|u\|_{L^\infty(B_R)} \leq C\left(R^{-d/\rho}\|u\|_{L^\rho(2B_R)} + R^{2-\frac{d}{r}}\|g\|_{L^r(\Omega)} + R^{1-\frac{d}{2r}}\|G\|_{L^{2r}(\Omega)}\right).$$

With these preliminaries in place, let us specialize our attention to the operators which appear most prominently in our work: Suppose that M takes the form

$$Mu = -\mathrm{div}(\mathcal{A}\nabla u),$$

where we now assume that the matrix $\mathcal{A}$ has $\mathbb{Z}^d$-periodic entries, and we continue to assume that $\mathcal{A}$ satisfies the uniform ellipticity and smoothness conditions

$$\mathcal{A}(x) \succeq I_d/\lambda, \quad \text{over } \mathbb{T}^d \tag{6.55}$$

$$\max_{1\leq i,j\leq d} \|\mathcal{A}_{ij}\|_{\mathcal{C}^{1+\beta}(\mathbb{T}^d)} \leq \lambda. \tag{6.56}$$

By integration by parts, it is easy to see that M is self-adjoint with respect to the $L^2(\mathbb{T}^d)$ inner product, and thus, that its range contains only mean-zero functions. We may therefore view M as an operator mapping $H^2(\mathbb{T}^d)$ into $L_0^2(\mathbb{T}^d)$, and by convention, we will always restrict the domain of M to $H_0^2(\mathbb{T}^d)$.

Given $f \in L_0^2(\mathbb{T}^d)$, we will say that a map $u \in H_0^1(\mathbb{T}^d)$ is a weak solution to the PDE

$$Mu = f, \quad \text{over } \mathbb{T}^d \tag{6.57}$$

if for every $v \in H_0^1(\mathbb{T}^d)$, it holds that

$$\langle f, v\rangle_{L^2(\mathbb{T}^d)} = \langle u, v\rangle_{\mathcal{A}} := \int_{\mathbb{T}^d} \langle \mathcal{A}\nabla u, \nabla v\rangle d\mathcal{L}.$$

Under conditions (6.55–6.56), it is straightforward to see that $\langle\cdot,\cdot\rangle_{\mathcal{A}}$ defines an inner product on $H_0^1(\mathbb{T}^d)$, which is equivalent to the standard inner product $\langle\cdot,\cdot\rangle_{H^1(\mathbb{T}^d)}$. The definition of weak solution is motivated by the following observation: if a weak solution u admits a representative which lies in $\mathcal{C}_0^2(\mathbb{T}^d)$, then it follows by integration by parts that

$$\langle f, v\rangle_{L^2(\mathbb{T}^d)} = \langle u, v\rangle_{\mathcal{A}} = \langle Mu, v\rangle_{L^2(\mathbb{T}^d)}, \quad \text{for all } v \in H_0^1(\mathbb{T}^d),$$

which implies that the equation $Mu = f$ is solved in the classical sense over $\mathbb{T}^d$. Thus, when they exist, classical solutions coincide with weak solutions.

Since $H_0^1(\mathbb{T}^d)$, endowed with the inner product $\langle\cdot,\cdot\rangle_{\mathcal{A}}$, is a Hilbert space, and since the linear functional $\langle f,\cdot\rangle_{L^2(\mathbb{T}^d)} : H_0^1(\mathbb{T}^d) \to \mathbb{R}$ is bounded, it follows from the Riesz representation theorem that the PDE (6.57) admits a unique weak solution $u \in H_0^1(\mathbb{T}^d)$ for any given $f \in L_0^2(\mathbb{T}^d)$. The following standard result, which we prove for completeness, shows that weak solutions in fact lie in $H_0^2(\mathbb{T}^d)$, and are classically differentiable if $f \in \mathcal{C}_0^\beta(\mathbb{T}^d)$.

Lemma 77. *Assume that conditions* (6.55)–(6.56) *hold. Then, the operator M is a bijection of $H_0^2(\mathbb{T}^d)$ onto $L_0^2(\mathbb{T}^d)$, and its restriction to $\mathcal{C}_0^{2+\beta}(\mathbb{T}^d)$ is a bijection onto $\mathcal{C}_0^\beta(\mathbb{T}^d)$. Furthermore, it holds that*

$$\|Mu\|_{L^2(\mathbb{T}^d)} \asymp \|u\|_{H^2(\mathbb{T}^d)}, \quad \textit{for all } u \in H_0^2(\mathbb{T}^d),$$

and

$$\|Mu\|_{\mathcal{C}^\beta(\mathbb{T}^d)} \asymp \|u\|_{\mathcal{C}^{2+\beta}(\mathbb{T}^d)}, \quad \textit{for all } u \in \mathcal{C}_0^{2+\beta}(\mathbb{T}^d),$$

where the implicit constants depend only on λ, β, d.

Proof of Lemma 77. Let $f \in L_0^2(\mathbb{T}^d)$. By interior regularity estimates for weak solutions (Evans, 1998), the unique weak solution $u \in H_0^1(\mathbb{T}^d)$ to the equation $Mu = f$ over $\mathbb{T}^d$ in fact lies in $H_0^2(\mathbb{T}^d)$, and satisfies

$$\|u\|_{H^2(\mathbb{T}^d)} \leq C\Big(\|u\|_{L^2(\mathbb{T}^d)} + \|f\|_{L^2(\mathbb{T}^d)}\Big), \tag{6.58}$$

for a constant $C = C(\lambda, d, \beta) > 0$. The bijectivity of $M : H_0^2(\mathbb{T}^d) \to L_0^2(\mathbb{T}^d)$ immediately follows from here. Since the entries of $\mathcal{A}$ are bounded, M is is readily seen to be bounded, with operator norm bounded above by a constant depending only on λ, d. To bound the inverse, notice that for all u as before,

$$\|u\|_{L^2(\mathbb{T}^d)}^2 \leq \|u\|_{H^1(\mathbb{T}^d)}^2 \asymp \|u\|_A^2 = \langle u, f\rangle_{L^2(\mathbb{T}^d)} \leq \|u\|_{L^2(\mathbb{T}^d)}\|f\|_{L^2(\mathbb{T}^d)},$$

with implicit constant depending only on λ, d. We deduce that

$$\|u\|_{L^2(\mathbb{T}^d)} \lesssim \|f\|_{L^2(\mathbb{T}^d)}. \tag{6.59}$$

Combine this inequality with equation (6.58) to deduce that, for a possibly different constant $C > 0$,

$$\|u\|_{H^2(\mathbb{T}^d)} \leq C\|f\|_{L^2(\mathbb{T}^d)}.$$

This shows that M^{-1} has operator norm bounded again by a constant depending only on λ, d, β. It remains to prove the claims about Hölder continuity, for which we reason as in Theorem 6.8 of Gilbarg and Trudinger (2001). The injectivity of the operator $M : \mathcal{C}_0^{2+\beta}(\mathbb{T}^d) \to \mathcal{C}_0^{\beta}(\mathbb{T}^d)$ is a consequence of what we have already shown. To prove surjectivity, we use the method of continuity (see Gilbarg and Trudinger (2001), Section 5.2).

Lemma 78 (Method of Continuity)**.** *Let B_1 and B_2 be Banach spaces, and let M_0 and M_1 be bounded linear operators from B_1 into B_2. For all $s \in [0,1]$, let*

$$M_s = (1-s)M_0 + sM_1,$$

and assume there exists a constant $C > 0$ such that for any $s \in [0,1]$,

$$\|u\|_{B_1} \leq C\|M_s u\|_{B_2}. \tag{6.60}$$

Then, M_1 is a surjection of B_1 onto B_2 if and only if M_0 is a surjection of B_1 onto B_2.

We will apply the method of continuity with $M_0 = M$ and $M_1 = -\Delta$. Notice that M_s takes the form

$$M_s u = -\mathrm{div}(\mathcal{A}_s \nabla u), \quad \text{with } \mathcal{A}_s = (1-s)\mathcal{A} + sI_d,$$

and thus, up to modifying the value of λ, the matrix $\mathcal{A}_s$ satisfies conditions (6.55)–(6.56) uniformly in s. By what we have already shown, it follows that $M_s : H_0^2(\mathbb{T}^d) \to L_0^2(\mathbb{T}^d)$ is a bijection with $\|M_s u\|_{L^2(\mathbb{T}^d)} \asymp_{\lambda,\beta,d} \|u\|_{H^2(\mathbb{T}^d)}$. Thus, by Lemmas 25(i) and 76,

$$\|u\|_{\mathcal{C}^{2+\beta}(\mathbb{T}^d)} \lesssim_{\lambda,d,\beta} \|u\|_{L^2(\mathbb{T}^d)} + \|M_s u\|_{\mathcal{C}^{\beta}(\mathbb{T}^d)} \asymp \|M_s u\|_{\mathcal{C}^{\beta}(\mathbb{T}^d)},$$

and hence, by the method of continuity, M is a bijection, whose inverse has norm bounded by a constant depending only on λ, d, β. Finally, it is also easy to see that M has norm bounded above by such a constant, by Lemma 95. The claim follows. □

It will be convenient to further note that the operator M admits a periodic Green's function, and hence that solutions u of the equation $Mu = f$ over $\mathbb{T}^d$ are in fact integral operators applied to f. We say that a map $\Gamma : \mathbb{T}^d \times \mathbb{T}^d \to \mathbb{R}$ is a periodic Green's function for M if for all $x, y \in \mathbb{T}^d$, $\Gamma(\cdot, x) \in L_0^1(\mathbb{T}^d)$, $\Gamma(y, \cdot) \in L_0^1(\mathbb{T}^d)$, and

$$v(x) = \langle \Gamma(\cdot, x), Mv \rangle_{L^2(\mathbb{T}^d)}, \quad \text{for all } v \in \mathcal{C}_0^{\infty}(\mathbb{T}^d).$$

As an example, it can be deduced from Stein and Weiss (1971) that $-\Delta$ admits a periodic Green's function over $\mathbb{T}^d$ which takes the form

$$\Gamma(y, x) = \Gamma_0(y - x) + b(y, x),$$

where $b \in \mathcal{C}^{\infty}(\mathcal{Q})$, and Γ_0 is the traditional periodic Green's function of the Laplace equation on $\mathbb{R}^d$, that is,

$$\Gamma_0(y - x) = \frac{1}{d(2-d)\omega_d}\|x - y\|^{2-d} \quad (d \geq 3),$$

where ω_d is the volume of the unit ball in dimension d. The following result shows that the operator M also admits a periodic Green's function, which has the same blow-up behaviour as Γ_0.

Proposition 39. Assume that conditions (6.55)–(6.56) hold, and let $d \geq 3$. Then, there exists a unique periodic Green's function Γ for the operator M, satisfying $\Gamma(\cdot, x) \in \mathcal{C}^1_{\mathrm{loc}}(\mathcal{Q} \setminus \{x\})$ for all $x \in \mathcal{Q}$, and for which there exists a constant $C = C(\lambda, \beta, d) > 0$ such that

$$|\Gamma(y,x)| \leq C\|x-y\|_{\mathbb{T}^d}^{2-d}, \quad \|\nabla\Gamma(y,x)\| \leq C\|x-y\|_{\mathbb{T}^d}^{1-d},$$

for all $x, y \in \mathbb{T}^d$, $x \neq y$. Furthermore, it holds that $\Gamma(x,y) = \Gamma(y,x)$ for any $x, y \in \mathbb{T}^d$. Finally, for any $f \in \mathcal{C}_0^{\beta}(\mathbb{T}^d)$ and $x \in \mathbb{T}^d$, it holds that

$$M^{-1}f(x) = \int_{\mathbb{T}^d} \Gamma(y,x)f(y)dy, \quad \nabla M^{-1}f(x) = \int_{\mathbb{T}^d} \nabla_y \Gamma(y;x)f(y)dy.$$

For uniformly elliptic operators over $\mathbb{R}^d$, analogues of Proposition 39 are classical; see for instance Littman, Stampacchia, and Weinberger (1963), Stampacchia (1965), and Grüter and Widman (1982). Over the torus $\mathbb{T}^d$, Proposition 39 can be deduced from Josien (2019).

6.B Additional Proofs from Section 6.3

6.B.1 Proof of Lemma 60

Given $u \in H_0^2(\mathbb{T}^d)$, let $f = Lu$. By Lemma 59, u is a weak $H_0^1(\mathbb{T}^d)$ solution to the equation

$$Eu = f(\nabla\varphi_0^*)\det(\nabla\varphi_0^*), \quad \text{over } \mathbb{T}^d,$$

and hence satisfies

$$\langle u, v\rangle_A = \langle f(\nabla\varphi_0^*)\det(\nabla\varphi_0^*), v\rangle_{L^2(\mathbb{T}^d)} = \langle f, v(\nabla\varphi_0^*)\rangle_{L^2(\mathbb{T}^d)},$$

for all $v \in H_0^1(\mathbb{T}^d)$. Using the fact that $\langle\cdot,\cdot\rangle_A$ defines an equivalent inner product to $\langle\cdot,\cdot\rangle_{H^1(\mathbb{T}^d)}$, we deduce that

$$\|u\|^2_{H^1(\mathbb{T}^d)} \lesssim \langle u,u\rangle_A = \langle f, u(\nabla\varphi_0^*)\rangle_{L^2(\mathbb{T}^d)} \leq \|f\|_{H^{-1}(\mathbb{T}^d)}\|u(\nabla\varphi_0^*)\|_{H^1(\mathbb{T}^d)} \lesssim \|f\|_{H^{-1}(\mathbb{T}^d)}\|u\|_{H^1(\mathbb{T}^d)},$$

where we used Lemma 82. We thus have $\|u\|_{H^1(\mathbb{T}^d)} \lesssim \|f\|_{H^{-1}(\mathbb{T}^d)}$, and to prove the reverse inequality, use Lemma 97 to write

$$\begin{aligned} \|f\|_{H^{-1}(\mathbb{T}^d)} &\asymp \sup_{\substack{v \in H^1(\mathbb{T}^d) \\ \|v\|_{H^1(\mathbb{T}^d)}=1}} \langle f, v\rangle_{L^2(\mathbb{T}^d)} \\ &= \sup_{\substack{v \in H^1(\mathbb{T}^d) \\ \|v\|_{H^1(\mathbb{T}^d)}=1}} \langle u, v(\nabla\varphi_0)\rangle_A \leq \|u\|_{H^1(\mathbb{T}^d)} \sup_{\substack{v \in H^1(\mathbb{T}^d) \\ \|v\|_{H^1(\mathbb{T}^d)}=1}} \|v(\nabla\varphi_0)\|_{H^1(\mathbb{T}^d)}. \end{aligned}$$

The claim now follows from Lemma 82. □

6.B.2 Proof of Lemma 61

We will make use of the following Lemma, which is proven below. In what follows, for any integers $d, k \geq 1$ and any differentiable function $f : \mathbb{R}^{d\times k} \to \mathbb{R}$, we denote by $\partial f(A)$ the $d \times k$ matrix with entries $(\partial f/\partial x_{ij})(A)$, $i = 1, \dots, d$, $j = 1, \dots, k$. Furthermore, for any open subset $\Theta \subseteq \mathbb{R}^{d\times k}$ and $\alpha > 0$, we denote by $\mathcal{C}^\alpha(\Omega; \Theta)$ the set of matrix-valued maps $A : \mathcal{Q} \to \Theta$ with entries in $\mathcal{C}^\alpha(\mathcal{Q})$.

Lemma 79. *Let $d, k \geq 1$ be integers, and $\Theta \subseteq \mathbb{R}^{d\times k}$. Let $f \in \mathcal{C}^2(\Theta)$ and $A, B \in \mathcal{C}^1(\mathcal{Q}; \Theta)$. Then, there exists a universal constant $C > 0$ such that function*

$$g : \mathcal{Q} \to \mathbb{R}, \quad g(x) = f(A(x) + B(x)) - f(A(x)) - \langle \partial f(A(x)), B(x)\rangle$$

satisfies

$$\|g\|_{\mathcal{C}^\beta(\mathcal{Q})} \leq C\|f\|_{\mathcal{C}^{2+\beta}(\Theta)}\Big(1 \vee \|A\|^\beta_{\mathcal{C}^1(\mathcal{Q};\Theta)}\Big)\|B\|^2_{\mathcal{C}^1(\mathcal{Q};\Theta)}.$$

A proof of Lemma 79 appears in Appendix 6.B.3. We now turn to the claim. We prove the Fréchet differentiability of $\widehat{\Psi}$, and a symmetric argument may be used for Ψ. Let $u \in \mathcal{C}_0^{2+\beta}(\mathcal{Q})$. Then,

$$\begin{aligned}
&\big\|\widehat{\Psi}[\varphi + u] - \widehat{\Psi}[\varphi] - \widehat{\Psi}'_\varphi[u]\big\|_{\mathcal{C}^\beta(\mathcal{Q})} \\
&\quad = \big\| \det(\nabla^2\varphi)\widehat{q}(\nabla\varphi) - \det(\nabla^2\varphi + \nabla^2 u)\widehat{q}(\nabla\varphi + \nabla u) \\
&\qquad\quad + \det(\nabla^2\varphi)\big[\widehat{q}(\nabla\varphi)\langle(\nabla^2\varphi)^{-1}, \nabla^2 u\rangle + \langle\nabla u, \nabla\widehat{q}(\nabla\varphi)\rangle\big]\big\|_{\mathcal{C}^\beta(\mathcal{Q})} \\
&\quad \leq \big\| \det(\nabla^2\varphi)\,[\widehat{q}(\nabla\varphi) - \widehat{q}(\nabla\varphi+\nabla u) - \langle\nabla u, \nabla\widehat{q}(\nabla\varphi)\rangle]\,\big\|_{\mathcal{C}^\beta(\mathcal{Q})} \\
&\quad + \big\|\widehat{q}(\nabla\varphi+\nabla u)\,\big[\det(\nabla^2\varphi) - \det(\nabla^2\varphi+\nabla^2 u)\big] + \widehat{q}(\nabla\varphi)\det(\nabla^2\varphi)\langle(\nabla^2\varphi)^{-1}, \nabla^2 u\rangle\big\|_{\mathcal{C}^\beta(\mathcal{Q})} \\
&\quad =: (I) + (II).
\end{aligned}$$

We first bound (I). By Lemma 95, we have,

$$(I) \lesssim \|\varphi\|_{\mathcal{C}^{2+\beta}(\mathcal{Q})}\big\|\widehat{q}(\nabla\varphi) - \widehat{q}(\nabla\varphi+\nabla u) - \langle\nabla u, \nabla\widehat{q}(\nabla\varphi)\rangle\big\|_{\mathcal{C}^\beta(\mathcal{Q})}.$$

Now, since $\widehat{q} \in \mathcal{C}^{2+\beta}(\mathcal{Q})$, we may invoke Lemma 79 to deduce

$$(I) \lesssim \|\widehat{q}\|_{\mathcal{C}^{2+\beta}(\mathcal{Q})}(1 + \|\varphi\|^{1+\beta}_{\mathcal{C}^{2+\beta}(\mathcal{Q})})\|u\|^2_{\mathcal{C}^1(\mathcal{Q})} \lesssim \|u\|^2_{\mathcal{C}^{2+\beta}(\mathcal{Q})}.$$

To bound term (II), apply Lemma 95 to obtain, almost surely,

$$\begin{aligned}
(II) &= \big\|\widehat{q}(\nabla\varphi+\nabla u)\,\big[\det(\nabla^2\varphi) - \det(\nabla^2\varphi+\nabla^2 u)\big] + \widehat{q}(\nabla\varphi)\det(\nabla^2\varphi)\langle(\nabla^2\varphi)^{-1}, \nabla^2 u\rangle\big\|_{\mathcal{C}^\beta(\mathcal{Q})} \\
&\leq \big\|\widehat{q}(\nabla\varphi+\nabla u)\,\big[\det(\nabla^2\varphi) - \det(\nabla^2\varphi+\nabla^2 u) + \det(\nabla^2\varphi)\langle(\nabla^2\varphi)^{-1}, \nabla^2 u\rangle\big]\,\big\|_{\mathcal{C}^\beta(\mathcal{Q})} \\
&+ \big\|\big[\widehat{q}(\nabla\varphi + \nabla u) - \widehat{q}(\nabla\varphi)\big]\det(\nabla^2\varphi)\langle(\nabla^2\varphi)^{-1}, \nabla^2 u\rangle\big\|_{\mathcal{C}^\beta(\mathcal{Q})} \\
&\lesssim \big\|\det(\nabla^2\varphi) - \det(\nabla^2\varphi+\nabla^2 u) + \det(\nabla^2\varphi)\langle(\nabla^2\varphi)^{-1}, \nabla^2 u\rangle\big\|_{\mathcal{C}^\beta(\mathcal{Q})} \\
&+ \|u\|_{\mathcal{C}^{2+\beta}(\mathbb{T}^d)}\big\|\widehat{q}(\nabla\varphi + \nabla u) - \widehat{q}(\nabla\varphi)\big\|_{\mathcal{C}^\beta(\mathcal{Q})}.
\end{aligned}$$

By Lemma A.1 of Figalli (2017) and Lemma 79, the first term in the final line of the above display is of order $O(\|u\|^2_{\mathcal{C}^{2+\beta}(\mathcal{Q})})$. Furthermore, since $\widehat{q} \in \mathcal{C}^1(\mathcal{Q})$, the second term is of order $O(\|u\|^2_{\mathcal{C}^{2+\beta}(\mathcal{Q})})$. The claim follows. □

6.B.3 Proof of Lemma 79

By a first-order Taylor expansion, it readily holds that for all $x \in \mathcal{Q}$,

$$|g(x)| \lesssim \|f\|_{\mathcal{C}^2(\Theta)}\|B(x)\|^2 \lesssim \|f\|_{\mathcal{C}^2(\Theta)}\|B\|^2_{\mathcal{C}^1(\Theta)}$$

It thus suffices to show that g is uniformly β-Hölder continuous, i.e. $|g(x)-g(y)| \leq L\|x-y\|^\beta$ for some $L > 0$ and all $x, y \in \mathcal{Q}$. To do so, define for all such x, y the function

$$h(x,y) = f(A(x)+B(y)) - f(A(x)) - \langle \partial f(A(x)), B(y)\rangle.$$

We have,

$$\begin{aligned}|g(x) - h(x,y)| &= |f(A(x)+B(x)) - f(A(x)+B(y)) - \langle \partial f(A(x)), B(x)-B(y)\rangle| \\ &\leq |f(A(x)+B(x)) - f(A(x)+B(y)) - \langle \partial f(A(x)+B(y)), B(x)-B(y)\rangle| \\ &+ \|\partial f(A(x)+B(y)) - \partial f(A(x))\|\|B(x)-B(y)\| =: (a)+(b).\end{aligned}$$

Clearly,

$$(a) \leq \|f\|_{\mathcal{C}^2(\Theta)}\|B(x)-B(y)\|^2 \leq \|f\|_{\mathcal{C}^2(\Theta)}\|B\|^2_{\mathcal{C}^1(\mathcal{Q},\Theta)}\|x-y\|^2.$$

Furthermore,

$$(b) \leq \|f\|_{\mathcal{C}^2(\Theta)}\|B(y)\|\|B(x)-B(y)\| \leq \|f\|_{\mathcal{C}^2(\Theta)}\|B\|^2_{\mathcal{C}^1(\mathcal{Q})}\|x-y\|.$$

Thus,

$$|g(x) - h(x,y)| \leq \|f\|_{\mathcal{C}^2(\Theta)}\|B\|^2_{\mathcal{C}^1(\mathcal{Q})}\|x-y\|.$$

Furthermore, we have

$$\begin{aligned}\big|h(x,y) - g(y)\big| &= \Big|\big[f(A(x)+B(y)) - f(A(y)+B(y))\big] + \big[f(A(x)) - f(A(y))\big] \\ &\qquad + \langle \partial f(A(x)) - \partial f(A(y)), B(y)\rangle\Big| \\ &\leq \sum_{i=1}^d \sum_{j=1}^k \left|\frac{\partial^2 f}{\partial X_{ij}}(A(x)) - \frac{\partial^2 f}{\partial X_{ij}}(A(y))\right| |B_i(y)B_j(y)| \\ &\leq \|f\|_{\mathcal{C}^{2+\beta}(\Theta)}\|B\|^2_{L^\infty(\Theta)}\|A(x)-A(y)\|^\beta \\ &\leq \|f\|_{\mathcal{C}^{2+\beta}(\Theta)}\|B\|^2_{\mathcal{C}^1(\Theta)}\|A\|^\beta_{\mathcal{C}^1(\Theta)}\|x-y\|^\beta.\end{aligned}$$

The claim follows from here. □

6.B.4 Proof of Lemma 62

By Lemma 61,

$$\big\|\widehat{\Psi}'_{\varphi_0}[u] - \Psi'_{\varphi_0}[u]\big\|_{\mathcal{C}^\beta(\mathbb{T}^d)}$$

$$
\begin{aligned}
&= \left\| \det(\nabla^2\varphi_0)\Big[\big(\widehat{q}(\nabla\varphi_0) - q(\nabla\varphi_0)\big)\langle\nabla^2\varphi_0, \nabla^2 u\rangle + \langle \nabla u, \nabla\widehat{q}(\nabla\varphi_0) - \nabla q(\nabla\varphi_0)\rangle\Big]\right\|_{\mathcal{C}^\beta(\mathbb{T}^d)} \\
&\lesssim \|\widehat{q}(\nabla\varphi_0) - q(\nabla\varphi_0)\|_{\mathcal{C}^\beta(\mathbb{T}^d)}\|\nabla^2 u\|_{\mathcal{C}^\beta(\mathbb{T}^d)} + \|\nabla u\|_{\mathcal{C}^\beta(\mathbb{T}^d)}\|\nabla\widehat{q}(\nabla\varphi_0) - \nabla q(\nabla\varphi_0)\|_{\mathcal{C}^\beta(\mathbb{T}^d)},
\end{aligned}
$$

where we used Lemma 95 to bound the Hölder norms of products. By Lemma 96, we further deduce,

$$
\begin{aligned}
\big\|\widehat{\Psi}'_{\varphi_0}[u] - \Psi'_{\varphi_0}[u]\big\|_{\mathcal{C}^\beta(\mathbb{T}^d)} &\lesssim \|\widehat{q} - q\|_{\mathcal{C}^\beta(\mathbb{T}^d)}\|\nabla^2 u\|_{\mathcal{C}^\beta(\mathbb{T}^d)} + \|\nabla u\|_{\mathcal{C}^\beta(\mathbb{T}^d)}\|\nabla\widehat{q} - \nabla q\|_{\mathcal{C}^\beta(\mathbb{T}^d)} \\
&\lesssim \|u\|_{\mathcal{C}^{2+\beta}(\mathbb{T}^d)}\|\widehat{q} - q\|_{\mathcal{C}^\beta(\mathbb{T}^d)} + \|u\|_{\mathcal{C}^{1+\beta}(\mathbb{T}^d)}\|\widehat{q} - q\|_{\mathcal{C}^{1+\beta}(\mathbb{T}^d)},
\end{aligned}
$$

as was to be shown. □

6.C Additional Proofs from Section 6.4

6.C.1 Proof of Lemma 65

Let $u = E_0^{-1}[f(S_{Y,x}(\cdot))]$. Notice that $\mathcal{B}_0$ is symmetric, thus we have for all $y \in \mathbb{T}^d$,

$$
\begin{aligned}
f(y) &= E_0[u](S_{Y,x}^{-1}(y)) \\
&= -\mathrm{tr}\big(A_0^\top \nabla^2 u(S_{Y,x}^{-1}(y))\big) \\
&= -\mathrm{tr}\big((\mathcal{B}_0^2 A_0)^\top \mathcal{B}_0^{-2}\nabla^2 u(S_{Y,x}^{-1}(y))\big) \\
&= -\mathrm{tr}\big((\mathcal{B}_0^2 A_0)^\top \nabla_y^2 u(S_{Y,x}^{-1}(y))\big) \\
&= G_0[u(S_{Y,x}^{-1}(\cdot))](y)
\end{aligned}
$$

It follows that $u = G_0^{-1}[f](S_{Y,x}(y)) = E_0^{-1}[f(S_{Y,x}(\cdot))](y)$, as was to be shown. □

6.C.2 Proof of Lemma 66

To prove the claim, it suffices to show that

$$
\left\|\mathrm{Cov}\big\{\nabla G_0^{-1}[\overline{K}_{h_n}^o](Y - \nabla\varphi_0(x))\big\} - q(\nabla\varphi_0(x))\mathrm{Cov}\big\{\nabla G_0^{-1}[\overline{K}_{h_n}^o](U)\big\}\right\| \lesssim h_n^{2+\epsilon-d}
$$

for some $\epsilon > 0$. Let $R(u) = \nabla G_0^{-1}[\overline{K}_{h_n}^o](u)$, and notice that

$$
\begin{aligned}
&\left\|\mathbb{E}[R(Y - \nabla\varphi_0(x))R(Y - \nabla\varphi_0(x))^\top] - q(\nabla\varphi_0(x))\mathbb{E}[R(U)R(U)^\top]\right\| \\
&\quad= \left\|\int_{\mathbb{T}^d} R(y)R(y)^\top q(y + \nabla\varphi_0(x))dy - q(\nabla\varphi_0(x))\int_{\mathbb{T}^d} R(y)R(y)^\top dy\right| \\
&\quad= \left\|\int_{\mathbb{T}^d} R(y)R(y)^\top [q(y + \nabla\varphi_0(x)) - q(\nabla\varphi_0(x))]\right\| \\
&\quad\lesssim \int_{\mathbb{T}^d} \|R(y)R(y)^\top\| \|y\|_{\mathbb{T}^d} dy \\
&\quad\lesssim \int_{\mathbb{T}^d} \|R(y)\|^2 \|y\|_{\mathbb{T}^d} dy,
\end{aligned}
$$

where we used the fact that $q \in \mathcal{C}^1(\mathbb{T}^d)$. Reasoning in the same way as we did below the statement of Lemma 66, we have $\mathbb{E}[R(U)] = 0$, and thus

$$\Big\|\mathbb{E}[R(Y - \nabla\varphi_0(x))]\Big\| = \Big\|\mathbb{E}[R(Y - \nabla\varphi_0(x))] - q(\nabla\varphi_0(x))\mathbb{E}[R(U)]\Big\| \lesssim \int_{\mathbb{T}^d} \|R(y)\|\|y\|_{\mathbb{T}^d} dy.$$

By combining the previous two displays, and using Jensen's inequality, we have

$$\Big\|\mathrm{Cov}\big\{R(Y - \nabla\varphi_0(x))\big\} - q(\nabla\varphi_0(x))\mathrm{Cov}\big\{R(U)\big\}\Big\| \lesssim \mathcal{I} := \int_{\mathbb{T}^d} \|R(y)\|^2\|y\|dy,$$

and it thus remains to bound $\mathcal{I}$. To this end, recall from Proposition 39 that G_0 admits a periodic Green's function $\Gamma_{G_0}(y,z)$, which is continuously differentiable away from the diagonal $z = y$, and whose gradient satisfies $\|\nabla\Gamma_{G_0}(y,z)\| \lesssim \|y-z\|_{\mathbb{T}^d}^{1-d}$. Deduce that,

$$\begin{aligned}
\mathcal{I} &\lesssim \int_{\mathbb{T}^d} \left(\int_{\mathbb{T}^d} \|y-z\|_{\mathbb{T}^d}^{1-d}\big|\overline{K}^o_{h_n}(z)\big|dz\right)^2 \|y\|_{\mathbb{T}^d} dy \\
&\lesssim h_n^{-2d} \int_{[-1/2,1/2]^d} \left(\int_{B(0,h_n)} \|y-z\|_{\mathbb{T}^d}^{1-d} dz\right)^2 \|y\|dy \\
&\lesssim h_n^{-2d} \int_{[-1/2,1/2]^d} \int_{B(0,h_n)} \int_{B(0,h_n)} \|y-z\|_{\mathbb{T}^d}^{1-d}\|y-w\|_{\mathbb{T}^d}^{1-d}\|y\|dzdwdy \\
&\lesssim 1 + \mathcal{I}_1 + \mathcal{I}_2,
\end{aligned}$$

where, given a constant $a > 2$ satisfying $(2-2d)/a \geq (5/2) - d$, we define

$$\begin{aligned}
\mathcal{I}_1 &= h_n^{-2d} \int_{B(0,h_n^{1/a})} \int_{B(0,h_n)} \int_{B(0,h_n)} \|y-z\|_{\mathbb{T}^d}^{1-d}\|y-w\|_{\mathbb{T}^d}^{1-d}\|y\|dzdwdy \\
\mathcal{I}_2 &= h_n^{-2d} \int_{B(0,1/4)\setminus B(0,h_n^{1/a})} \int_{B(0,h_n)} \int_{B(0,h_n)} \|y-z\|_{\mathbb{T}^d}^{1-d}\|y-w\|_{\mathbb{T}^d}^{1-d}\|y\|dzdwdy.
\end{aligned}$$

Notice that in both of the above integrations, we in fact have $\|y-z\|_{\mathbb{T}^d} = \|y-z\|$ and $\|y-w\|_{\mathbb{T}^d} = \|y-w\|$. To bound $\mathcal{I}_1$, invoke Lemmas 83–84 to obtain

$$\begin{aligned}
\mathcal{I}_1 &\lesssim h_n^{\frac{1}{a}-2d} \int_{B(0,h_n)} \int_{B(0,h_n)} \|z-w\|^{2-d} dzdw \\
&\lesssim h_n^{\frac{1}{a}+2-2d} \int_{B(0,h_n)} dw \lesssim h_n^{\frac{1}{a}+2-d} \lesssim h_n^{\frac{5}{2}-d}.
\end{aligned}$$

Regarding $\mathcal{I}_2$, we have the crude bound

$$\mathcal{I}_2 = h_n^{-2d} \int_{B(0,1/4)\setminus B(0,h_n^{1/a})} \int_{B(0,h_n)} \int_{B(0,h_n)} \|y-z\|^{1-d}\|y-w\|^{1-d} dzdwdy \lesssim h_n^{\frac{2-2d}{a}}.$$

Combining these facts together with the definition of a, we arrive at the claim. □

6.C.3 Proof of Lemma 67

Abbreviate $f(\xi) = (\mathcal{F}[K](\xi)/2\pi)^2$. Define

$$g_n = h_n^{d-2} \sum_{\xi \in \mathbb{Z}^d_*} \xi\xi^\top \frac{f(h_n\xi)}{\langle \mathcal{C}_0\xi, \xi\rangle^2}, \quad g_\infty = \int_{\mathbb{R}^d} zz^\top \frac{f(z)dz}{\langle \mathcal{C}_0 z, z\rangle^2},$$

and for any $\epsilon > 0$, define

$$g_{n,\epsilon} = h_n^{d-2} \sum_{\substack{\xi \in \mathbb{Z}^d_* \\ \epsilon \le \|h_n\xi\|_\infty < \epsilon^{-1}}} \xi\xi^\top \frac{f(h_n\xi)}{\langle \mathcal{C}_0\xi, \xi\rangle^2}, \quad g_{\infty,\epsilon} = \int_{[\epsilon^{-1},\epsilon]^d} zz^\top \frac{f(z)dz}{\langle \mathcal{C}_0 z, z\rangle^2}.$$

Notice that

$$\begin{aligned} &g_{n,\epsilon} - g_{\infty,\epsilon} \\ &= \sum_{\substack{\xi \in \mathbb{Z}^d_* \\ \epsilon \le \|h_n\xi\|_\infty < \epsilon^{-1}}} \int_{I_{\xi,n}} \left[(h_n\xi)(h_n\xi)^\top \frac{f(h_n\xi)}{\langle \mathcal{C}_0(h_n\xi), (h_n\xi)\rangle^2} - zz^\top \frac{f(z)}{\langle \mathcal{C}_0 z, z\rangle^2} \right] dz \end{aligned}$$

where $I_{\xi,n}$ is the hypercube of side length h_n, with vertices lying in $h_n\mathbb{Z}^d$, whose vertex nearest to the origin is at the point $h_n\xi$. We will provide a very crude bound on the norm of the above quantity, which will suffice for our purposes. Notice that for all ξ satisfying $\epsilon \le \|h_n\xi\|_\infty < \epsilon^{-1}$ and $z \in I_{\xi,n}$, we have

$$\begin{aligned} &\left\| (h_n\xi)(h_n\xi)^\top \frac{f(h_n\xi)}{\langle \mathcal{C}_0(h_n\xi), (h_n\xi)\rangle^2} - zz^\top \frac{f(z)}{\langle \mathcal{C}_0 z, z\rangle^2} \right\| \\ &\quad \le \|(h_n\xi)(h_n\xi)^\top\| \left| \frac{f(h_n\xi)}{\langle \mathcal{C}_0(h_n\xi), (h_n\xi)\rangle^2} - \frac{f(z)}{\langle \mathcal{C}_0 z, z\rangle^2} \right| \\ &\quad + \left| \frac{f(z)}{\langle \mathcal{C}_0 z, z\rangle^2} \right| \left\| (h_n\xi)(h_n\xi)^\top - zz^\top \right\| \\ &\quad \lesssim \epsilon^{-2} \left| \frac{f(h_n\xi)}{\langle \mathcal{C}_0(h_n\xi), (h_n\xi)\rangle^2} - \frac{f(z)}{\langle \mathcal{C}_0 z, z\rangle^2} \right| + \epsilon^{-5} h_n, \end{aligned}$$

where we used the fact that f is bounded in the final line, and the fact that $\mathrm{diam}(I_{\xi,n}) \lesssim h_n$. Furthermore,

$$\begin{aligned} &\left| \frac{f(h_n\xi)}{\langle \mathcal{C}_0(h_n\xi), (h_n\xi)\rangle^2} - \frac{f(z)}{\langle \mathcal{C}_0 z, z\rangle^2} \right| \\ &\le |f(h_n\xi)| \left| \frac{1}{\langle \mathcal{C}_0(h_n\xi), (h_n\xi)\rangle^2} - \frac{1}{\langle \mathcal{C}_0 z, z\rangle^2} \right| + \frac{1}{\langle \mathcal{C}_0 z, z\rangle^2} |f(h_n\xi) - f(z)| \\ &\lesssim \epsilon^{-8} \left| \langle \mathcal{C}_0(h_n\xi), (h_n\xi)\rangle^2 - \langle \mathcal{C}_0 z, z\rangle^2 \right| + \epsilon^{-4} h_n \\ &\lesssim \epsilon^{-10} \left| \langle \mathcal{C}_0(h_n\xi), (h_n\xi)\rangle - \langle \mathcal{C}_0 z, z\rangle \right| + \epsilon^{-4} h_n \\ &\lesssim \epsilon^{-11} \|h_n\xi - z\| + \epsilon^{-4} h_n \end{aligned}$$

$$\lesssim \epsilon^{-11} h_n + \epsilon^{-4} h_n.$$

Altogether, we have shown that there exists $a > 0$ such that

$$\|g_{n,\epsilon} - g_{\infty,\epsilon}\| \lesssim \epsilon^{-a} h_n. \tag{6.61}$$

Next, we bound the distance between $g_{n,\epsilon}$ and g_n. Since f is a Schwartz function, we have $|f(\xi)| \lesssim_\ell \|\xi\|^{-\ell}$ for all $\ell \geq 0$. Taking $\ell = d-1$, we obtain

$$\begin{aligned}
\left\| h_n^{d-2} \sum_{\substack{\xi \in \mathbb{Z}_*^d \\ \|\xi h_n\|_\infty \geq 1/\epsilon}} \xi\xi^\top \frac{f(h_n\xi)}{\langle \mathcal{C}_0 \xi, \xi\rangle^2} \right\| &\lesssim h_n^{d-2} \sum_{\substack{\xi \in \mathbb{Z}_*^d \\ \|\xi h_n\|_\infty \geq 1/\epsilon}} \frac{\|h_n \xi\|^{1-d}}{\|\xi\|^2} \\
&\lesssim h_n^{-1} \sum_{\substack{\xi \in \mathbb{Z}_*^d \\ \|\xi h_n\|_\infty \geq 1/\epsilon}} \|\xi\|^{-(d+1)} \\
&\lesssim h_n^{-1} \int_{c_d(\epsilon h_n)^{-1} - 1}^\infty r^{-(d+1)} r^{d-1} dr \asymp \epsilon,
\end{aligned}$$

for a dimension-dependent constant $c_d > 0$. Similarly,

$$\begin{aligned}
\left\| h_n^{d-2} \sum_{\substack{\xi \in \mathbb{Z}_*^d \\ \|\xi h_n\|_\infty < \epsilon}} \xi\xi^\top \frac{f(h_n\xi)}{\langle \mathcal{C}_0 \xi, \xi\rangle^2} \right\| &\lesssim h_n^{d-2} \sum_{\substack{\xi \in \mathbb{Z}_*^d \\ \|\xi h_n\|_\infty < \epsilon}} \|\xi\|^{-2} \\
&\lesssim h_n^{d-2} \int_1^{\epsilon/h_n} r^{d-3} dr \asymp \epsilon^{d-2}.
\end{aligned}$$

Combining the preceding two displays, we thus have

$$\|g_{n,\epsilon} - g_n\| \lesssim \epsilon. \tag{6.62}$$

Analogous derivations show that

$$\|g_{\infty,\epsilon} - g_\infty\| \lesssim \epsilon. \tag{6.63}$$

Combining equations (6.61), (6.62), and (6.63), we deduce

$$\|g_n - g_\infty\| \lesssim \epsilon + \epsilon^{-a} h_n,$$

for a large enough constant $a > 0$. The claim now follows by taking $\epsilon \asymp h_n^{\frac{1}{1+a}}$. □

6.C.4 Proof of Lemma 68

Recall that $\mathrm{supp}(K_{h_n}) \subseteq B(0,h_n)$. Therefore, a point $y \in \mathbb{T}^d$ can only lie in the support of $K_{h_n}(Y - \nabla\varphi_0(\cdot))$ if

$$Y - \nabla\varphi_0(y) \in B(0,h_n), \text{ or equivalently, } y \in \nabla\varphi_0^*(B(Y,h_n)).$$

Since $\nabla\varphi_0^*$ is λ-Lipschitz, we have

$$\nabla\varphi_0^*(B(Y,h_n)) \subseteq B(\nabla\varphi_0^*(Y), \lambda h_n),$$

and we deduce that

$$\mathrm{supp}\left(K_{h_n}(Y - \nabla\varphi_0(\cdot))\right) \subseteq B(\nabla\varphi_0^*(Y), \lambda h_n).$$

Notice that S_x^{-1} is also Lipschitz with parameter λ, thus the same argument implies

$$\mathrm{supp}\left(K_{h_n}(Y - S_x(\cdot))\right) \subseteq B(S_x^{-1}(Y), \lambda h_n).$$

The claim follows. □

6.C.5 Proof of Lemma 69

Let $I = \mathbb{E}\left[\left(\int_{B(Z,\epsilon)} \|x-z\|_{\mathbb{T}^d}^{t-d} dz\right)^2\right]$. It holds that,

$$\begin{aligned}
I &\le \gamma \mathbb{E}_{Z\sim\mathcal{L}}\left[\left(\int_{B(Z,\epsilon)} \|x-z\|_{\mathbb{T}^d}^{t-d} dz\right)^2\right] \\
&= \int_{\mathbb{T}^d}\int_{B(y,\epsilon)}\int_{B(y,\epsilon)} \|x-z\|_{\mathbb{T}^d}^{t-d}\|x-w\|_{\mathbb{T}^d}^{t-d} dz dw dy \\
&= \int_{\mathbb{T}^d}\int_{B(0,\epsilon)}\int_{B(0,\epsilon)} \|x-z-y\|_{\mathbb{T}^d}^{t-d}\|x-w-y\|_{\mathbb{T}^d}^{t-d} dz dw dy \\
&= \int_{B(0,\epsilon)}\int_{B(0,\epsilon)} \left(\int_{\mathbb{T}^d} \|x-z-y\|_{\mathbb{T}^d}^{t-d}\|x-w-y\|_{\mathbb{T}^d}^{t-d} dy\right) dz dw \\
&= \int_{B(0,\epsilon)}\int_{B(0,\epsilon)} \left(\int_{\mathbb{T}^d} \|z-y\|_{\mathbb{T}^d}^{t-d}\|w-y\|_{\mathbb{T}^d}^{t-d} dy\right) dz dw \\
&\lesssim \int_{B(0,\epsilon)}\int_{B(0,\epsilon)} \|z-w\|_{\mathbb{T}^d}^{2t-d} dz dw \\
&\lesssim \int_{B(0,\epsilon)}\int_{B(0,\epsilon)} \|z-w\|^{2t-d} dz dw,
\end{aligned}$$

where the penultimate line follows from Lemma 83, together with the condition $t < d/2$. Passing to spherical coordinates, we obtain

$$I \le \int_{B(0,\epsilon)}\int_{B(0,\epsilon)} \big|\|z\| - \|w\|\big|^{2t-d} dz dw$$

$$\begin{aligned}
&\leq \int_0^\epsilon \int_0^\epsilon |r-s|^{2t-d} r^{d-1} s^{d-1} drds \\
&= \int_0^\epsilon \int_0^s (s-r)^{2t-d} r^{d-1} s^{d-1} drds + \int_0^\epsilon \int_s^\epsilon (r-s)^{2t-d} r^{d-1} s^{d-1} drds \\
&=: (A) + (B).
\end{aligned}$$

We bound terms (A) and (B) in turn. On the one hand,

$$(A) \leq \int_0^\epsilon \int_0^s s^{2t-d} r^{d-1} s^{d-1} drds = \int_0^\epsilon \int_0^s s^{2t-1} r^{d-1} drds \asymp \int_0^\epsilon s^{d+2t-1} ds \asymp \epsilon^{d+2t},$$

while on the other hand,

$$(B) \leq \int_0^\epsilon \int_s^\epsilon r^{2t-1} s^{d-1} drds \leq \int_0^\epsilon \epsilon^{2t} s^{d-1} ds \asymp \epsilon^{d+2t}.$$

This proves the claim. □

6.C.6 Proof of Lemma 70

By Lemmas 77, 95, and 96, we have

$$\|u_n\|_{\mathcal{C}^{2+\beta}(\mathbb{T}^d)} \lesssim \| \det(\mathcal{B}) \overline{K}^o_{h_n}(Y - \nabla\varphi_0(\cdot))\|_{\mathcal{C}^\beta(\mathbb{T}^d)} \lesssim \|\overline{K}^o_{h_n}\|_{\mathcal{C}^\beta(\mathbb{T}^d)} \lesssim h_n^{-(d+\beta)}.$$

Now, recall that $\lambda > 1$ is such that $\|\varphi_0\|_{\mathcal{C}^3(\mathbb{T}^d)} \leq \lambda$ and $\nabla^2 \varphi_0^* \succeq \lambda^{-1} I_d$ uniformly over $\mathbb{T}^d$. To prove the claim, it will be sufficient to show that for all $y \in \mathbb{T}^d$ such that $\|y - X_0\|_{\mathbb{T}^d} \geq 2\lambda h_n$, we have $\|\nabla^2 u_n(y)\| \lesssim \|X_0 - y\|_{\mathbb{T}^d}^{-d}$. To prove this, note that it suffices to consider all $y \in \mathbb{T}^d$ such that $\|X_0 - y\| = \|X_0 - y\|_{\mathbb{T}^d}$, and we shall assume this equality holds throughout the remainder of the proof.

Assume thus that $\|y - X_0\| \geq 2\lambda h_n$, and set

$$B = B(y, \|y - X_0\|/2), \quad B_0 = B(y, h_n/4).$$

Notice that for all $v \in B$,

$$\|v - X_0\| \geq \|y - X_0\| - \|v - y\| \geq \frac{\|y - X_0\|}{4}. \tag{6.64}$$

Furthermore, we have $\operatorname{dist}(B, B_0) \geq \|y - X_0\|/4$, thus we may apply the a priori bound in Lemma 25(i) to obtain

$$\|u_n\|_{\mathcal{C}^2(B_0)} \lesssim \|y - X_0\|^{-2} \Big[\| \det(\mathcal{B}) \overline{K}^o_{h_n}(Y - \nabla\varphi_0(\cdot))\|_{\mathcal{C}^\beta(B)} + \|u_n\|_{L^\infty(B)} \Big], \tag{6.65}$$

where we again used Lemma 95. Recall from Lemma 68 that $\operatorname{supp}(\overline{K}_{h_n}(Y - \nabla\varphi_0(\cdot))) \subseteq B(X_0, \lambda h_n)$. The latter is disjoint from B by equation (6.64), thus we have

$$\overline{K}^o_{h_n}(Y - \nabla\varphi_0(v)) = -1 \quad \text{for all } v \in B,$$

and hence, from equation (6.65) we obtain

$$\begin{aligned}\|u_n\|_{\mathcal{C}^2(B_0)} &\lesssim \|y-X_0\|^{-2}\Big[\|\det(\mathcal{B})\|_{\mathcal{C}^\beta(B)} + \|u_n\|_{L^\infty(B)}\Big] \\ &\lesssim \|y-X_0\|^{-2}\Big[1+\|u_n\|_{L^\infty(B)}\Big]. \end{aligned} \tag{6.66}$$

It thus remains to bound $\|u_n\|_{L^\infty(B)}$. To this end, recall that we can write for any $v \in B$,

$$u_n(v) = \int_{\mathbb{T}^d} \Gamma_E(z,v)\overline{K}^o_{h_n}(Y - \nabla\varphi_0(z))dz,$$

where Γ_E is the periodic Green's function of E (cf. Proposition 39). We thus have, for all $v \in B$,

$$\begin{aligned}|u_n(v)| &\lesssim \int_{\mathbb{T}^d} \|z-v\|_{\mathbb{T}^d}^{2-d}\big|\overline{K}^o_{h_n}(Y-\nabla\varphi_0(z))\big|dz \\ &\lesssim 1 + h_n^{-d}\int_{B(X_0,\lambda h_n)} \|z-v\|_{\mathbb{T}^d}^{2-d}dz \\ &= 1 + h_n^{-d}\int_{B(0,\lambda h_n)} \|z-X_0-v\|_{\mathbb{T}^d}^{2-d}dz \\ &\le 1 + h_n^{-d}\int_{B(0,\lambda h_n)} \big[\|X_0-v\|_{\mathbb{T}^d} - \lambda h_n\big]^{2-d}dz \\ &\lesssim 1 + \big[\|X_0-v\|_{\mathbb{T}^d} - \lambda h_n\big]^{2-d} \\ &\lesssim 1 + \|X_0-v\|_{\mathbb{T}^d}^{2-d}, \end{aligned} \tag{6.67}$$

where the final inequality follows from equation (6.64). From equations (6.66–6.67), we obtain

$$\begin{aligned}\|u_n\|_{\mathcal{C}^2(B_0)} &\lesssim \sup_{v\in B}\|y-X_0\|^{-2}\Big[1+\|X_0-v\|_{\mathbb{T}^d}^{2-d}\Big] \\ &\lesssim \sup_{v\in B}\|y-X_0\|^{-2}\|X_0-v\|_{\mathbb{T}^d}^{2-d} \\ &\le \|y-X_0\|^{-2}\big(\|X_0-y\|_{\mathbb{T}^d} - h_n/2\big)^{2-d} \lesssim \|y-X_0\|^{-d}. \end{aligned}$$

The claim follows. □

6.C.7 Proof of Lemma 71

We will make use of the following gradient estimate for u_n.

Lemma 80. *There exists a constant $C>0$ such that for all $y \in \mathbb{T}^d$,*

$$\|\nabla u_n(y)\| \le C\big(h_n \vee \|X_0-y\|_{\mathbb{T}^d}\big)^{1-d}.$$

The proof of Lemma 80 appears in Appendix 6.C.8. With this result in place, we can bound $\mathbb{E}[\mathcal{I}_2^2]$ in nearly the same way as $\mathbb{E}[\mathcal{I}_1^2]$. To see this, notice first that from Lemma 80, we have

$$\begin{aligned}\mathbb{E}&\left[\left(\int_{B(X_0,h_n)}\|x-y\|_{\mathbb{T}^d}^{1-d}\|\nabla u_n(y)\|dy\right)^2\right]\\ &\lesssim h_n^{2(1-d)}\mathbb{E}\left[\left(\int_{B(X_0,h_n)}\|x-y\|_{\mathbb{T}^d}^{1-d}dy\right)^2\right]\lesssim h_n^{4-d},\end{aligned}$$

by Lemma 69. Thus, as before,

$$\mathbb{E}[\mathcal{I}_2^2]\lesssim h_n^{4-d}+\mathbb{E}\left[\left(\int_{\mathcal{Q}_0\setminus B(0,h_n)}\|x-X_0-y\|_{\mathbb{T}^d}^{1-d}\|y\|_{\mathbb{T}^d}^{1-d}dy\right)^2\right].$$

Let the event A_n be defined as in Step 4 of the proof of Lemma 63, and recall equation (6.47), from which it follows that

$$\begin{aligned}\mathbb{E}&\left[\left(\int_{\mathcal{Q}_0\setminus B(0,h_n)}\|x-X_0-y\|_{\mathbb{T}^d}^{1-d}\|y\|_{\mathbb{T}^d}^{1-d}dy\right)^2 I(A_n)\right]\\ &\lesssim \mathbb{E}\left[\left(\int_{\mathcal{Q}_0\setminus B(0,h_n)}\|y\|^{2-2d}dy\right)^2 I(A_n)\right]\lesssim h_n^{4-d}.\end{aligned}$$

Furthermore, by Lemma 83,

$$\begin{aligned}\mathbb{E}&\left[\left(\int_{\mathcal{Q}_0\setminus B(0,h_n)}\|x-X_0-y\|_{\mathbb{T}^d}^{1-d}\|y\|^{1-d}dy\right)^2 I(A_n^c)\right]\\ &\lesssim \mathbb{E}\left[\left(\|x-X_0\|^{2-d}\right)^2 I(A_n^c)\right]\asymp h_n^{4-d}\vee 1.\end{aligned}$$

The claim follows. □

6.C.8 Proof of Lemma 80

From Proposition 39 and the definition of $\overline{K}_{h_n}^o$, we have for all $y\in\mathbb{T}^d$,

$$\|\nabla u_n(y)\|\lesssim 1+\int_{\mathbb{T}^d}\|y-z\|_{\mathbb{T}^d}^{1-d}\big|\overline{K}_{h_n}(Y-\nabla\varphi_0(z))\big|dz.$$

By Lemma 68, we have $\mathrm{supp}(\overline{K}_{h_n}(Y-\nabla\varphi_0(\cdot)))\subseteq B(X_0,\lambda h_n)$, thus

$$\|\nabla u_n(y)\|\lesssim 1+h_n^{-d}\int_{B(X_0,\lambda h_n)}\|y-z\|_{\mathbb{T}^d}^{1-d}dz\lesssim h_n^{1-d},$$

by Lemma 84. □

6.D Additional Proofs from Section 6.5

6.D.1 Proof of Lemma 72

Let $x \in \mathbb{T}^d$. Let us first show that $\mathbb{E}[\nabla Z_{n,i1}(x)] = 0$. Using Lemma 59 and Proposition 39, it can be deduced that

$$\nabla L^{-1}\big[\overline{K}_{h_n}(Y_1 - \cdot) - q_{h_n}\big](x) = \int_{\mathbb{T}^d} \nabla \Gamma_E(\nabla\varphi_0^*(y), x)\big[\overline{K}_{h_n}(Y_1 - y) - q_{h_n}(y)\big]dy,$$

where Γ_E is a periodic Green's function for the operator E, and so

$$\mathbb{E}[\nabla Z_{n,1}(x)] = \int_{\mathbb{T}^d}\int_{\mathbb{T}^d} \nabla \Gamma_E(\nabla\varphi_0^*(y), x)\big[\overline{K}_{h_n}(z - y) - q_{h_n}(y)\big]dydQ(z).$$

One may apply Fubini's theorem to change the order of integration in the above display, since for any given $h_n > 0$, by Tonelli's theorem,

$$\int_{\mathbb{T}^d}\int_{\mathbb{T}^d} \big\|\nabla \Gamma_E(\nabla\varphi_0^*(y), x)\big[\overline{K}_{h_n}(z - y) - q_{h_n}(y)\big]\big\|dydQ(z) \lesssim h_n^{-d}\int_{\mathbb{T}^d}\|\nabla\varphi_0^*(y) - x\|^{1-d}dy < \infty.$$

Deduce that

$$\mathbb{E}[\nabla Z_{n,i1}(x)] = \int_{\mathbb{T}^d} \nabla \Gamma_E(\nabla\varphi_0^*(y), x)\left(\int_{\mathbb{T}^d}\big[\overline{K}_{h_n}(z - y) - q_{h_n}(y)\big]dQ(z)\right)dy = 0.$$

From here, we deduce

$$\begin{aligned}&\|\mathbb{E}[((1/n)\textstyle\sum_{i=1}^n \nabla Z_{n,i}(x))\, I_{A_n}]\| \\ &\quad \le \|\mathbb{E}[((1/n)\textstyle\sum_{i=1}^n \nabla Z_{n,i}(x))]\| + \big\|\mathbb{E}[((1/n)\textstyle\sum_{i=1}^n \nabla Z_{n,i}(x))\, I_{A_n^{\mathsf{c}}}]\big\| \\ &\quad = \big\|\mathbb{E}[((1/n)\textstyle\sum_{i=1}^n \nabla Z_{n,i}(x))\, I_{A_n^{\mathsf{c}}}]\big\|.\end{aligned}$$

Now, recall from Lemma 59 that

$$\|Z_{n,1}\|_{\mathcal{C}^{2+\beta}(\mathbb{T}^d)} \asymp \|\overline{K}_{h_n}(Y_1 - \cdot) - q_{h_n}]\|_{\mathcal{C}^{\beta}(\mathbb{T}^d)} \lesssim h_n^{-(d+\beta)},$$

and we deduce that

$$\|\mathbb{E}[((1/n)\textstyle\sum_{i=1}^n \nabla Z_{n,i}(x))\, I_{A_n}]\| \lesssim h_n^{-(d+\beta)}\mathbb{P}(A_n^{\mathsf{c}}) \lesssim n^{a(d+\beta)-b}.$$

It is clear that the implicit constants in these various assertions do not depend on x, and the claim follows. □

6.D.2 Proof of Lemma 73

By Lemma 59, we have

$$\|\nabla Z_{n,1}(x)\| \lesssim \|L^{-1}[\widehat{q}_n - q_{h_n}]\|_{\mathcal{C}^{2+\beta}(\mathbb{T}^d)} \lesssim \|\widehat{q}_n - q_{h_n}\|_{\mathcal{C}^{\beta}(\mathbb{T}^d)} \lesssim h_n^{-(d+\beta)}.$$

Thus, using Lemma 63, we have

$$\mathbb{E}\|\nabla Z_{n,1}(x)\|^{2+\delta} \lesssim h_n^{-\delta(d+\beta)}\mathbb{E}\|\nabla Z_{n,1}(x)\|^2 \lesssim h_n^{-\delta(d+\beta)+2-d},$$

as claimed. □

6.D.3 Proof of Proposition 37

Let $u_n = L^{-1}[\widehat{q}_n - q]$. We will begin by bounding the quantity $\langle u_n, v\rangle_A$, for any $v \in H_0^1(\mathbb{T}^d)$. By Lemma 59, the function u_n solves the PDE

$$Eu_n = \det(\nabla^2\varphi_0)(\widehat{q}_n(\nabla\varphi_0) - q(\nabla\varphi_0)), \quad \text{over } \mathbb{T}^d,$$

in the classical sense, and hence also in the weak $H_0^1(\mathbb{T}^d)$ sense (cf. Appendix 6.A). Therefore, for all $v \in H_0^1(\mathbb{T}^d)$,

$$\begin{aligned}\langle u_n, v\rangle_A &= \langle \det(\nabla^2\varphi_0)(\widehat{q}_n(\nabla\varphi_0) - q(\nabla\varphi_0)), v\rangle_{L^2(\mathbb{T}^d)} \\ &= \langle \widehat{q}_n - q, v(\nabla\varphi_0^*)\rangle_{L^2(\mathbb{T}^d)} \\ &= \int_{\mathbb{T}^d} \big(v(\nabla\varphi_0^*) \star K_{h_n}\big) d(Q_n - Q) + \int_{\mathbb{T}^d} v(\nabla\varphi_0^*) d(Q_{h_n} - Q).\end{aligned}$$

Using a simple argument which will be presented in further detail in Lemma 89 below, it is easy to see that

$$\int_{\mathbb{T}^d} \big(v(\nabla\varphi_0^*) \star K_{h_n}\big) d(Q_n - Q) = O_p(n^{-1/2}),$$

and furthermore,

$$\left|\int_{\mathbb{T}^d} v(\nabla\varphi_0^*) d(Q_{h_n} - Q)\right| \leq \|v(\nabla\varphi_0^*)\|_{H^1(\mathbb{T}^d)} \|q_{h_n} - q\|_{H^{-1}(\mathbb{T}^d)} \lesssim \|q_{h_n} - q\|_{H^1(\mathbb{T}^d)},$$

where we used Lemma 82. By Proposition 43, the final factor decays on the order h_n^{s+1}, thus we have shown

$$\langle u_n, v\rangle_{L^2(\mathbb{T}^d)} = O_p(n^{-1/2} + h_n^{s+1}).$$

To prove the claim, it thus remains to quantify the discrepancy between $\langle u_n, v\rangle_A$ and $\langle \widehat{\varphi}_n - \varphi_0, v\rangle_A$. To this end, recall from equation (6.49) that, for any given $\beta > 0$ sufficiently small,

$$\big\|\nabla\widehat{\varphi}_n - \nabla\varphi_0 - \nabla L^{-1}[\widehat{q}_n - q]\big\|_{L^\infty(\mathbb{T}^d)} = O_p\left(h_n^{2s-1-3\beta} + \frac{1}{nh_n^{d+1+4\beta}}\right).$$

It readily follows that

$$\begin{aligned}\langle \widehat{\varphi}_n - \varphi_0, v\rangle_A &= \langle L^{-1}[\widehat{q}_n - q], v\rangle_A + O_p\left(h_n^{2s-1-3\beta} + \frac{1}{nh_n^{d+1+4\beta}}\right) \\ &= O_p\left(n^{-1/2} + h_n^{s+1} + \frac{1}{nh_n^{d+1+4\beta}}\right).\end{aligned}$$

The claim follows. □

6.E Additional Technical Lemmas

The following $L^r(\mathbb{T}^d)$ interpolation inequality is a consequence of the periodic Riesz-Thorin theorem (cf. Remark 3.6.1/4 of Schmeisser and Triebel (1987)).

Lemma 81 (Interpolation Inequality for $\boldsymbol{L^r(\mathbb{T}^d)}$ Norms)**.** *Let $2 < r < \infty$. Then, for all $\epsilon > 0$, there exists $C(\epsilon, d, r) > 0$ such that for all $f \in L^\infty(\mathbb{T}^d)$,*

$$\|f\|_{L^r(\mathbb{T}^d)} \leq C(\epsilon)\|f\|_{L^2(\mathbb{T}^d)} + \epsilon\|f\|_{L^\infty(\mathbb{T}^d)}.$$

We next prove a simple chain rule-type result for fractional Sobolev spaces.

Lemma 82 (Composition of Sobolev Functions)**.** *Let $r > 1$. Let $f \in \mathcal{C}^2(\mathbb{T}^d)$ be such that*

$$\nabla^2 f \succeq I_d/\lambda \text{ over } \mathbb{T}^d, \quad \text{and} \quad \|f\|_{\mathcal{C}^2(\mathbb{T}^d)} \leq \lambda,$$

for some $\lambda > 0$. Then, there exists a constant $C = C(\lambda, d, r) > 0$ such that for all $\alpha \in [0,1]$ and all $g \in H^{\alpha,r}(\mathbb{T}^d)$,

$$\|g \circ \nabla f\|_{H^{\alpha,r}(\mathbb{T}^d)} \lesssim C\|g\|_{H^{\alpha,r}(\mathbb{T}^d)}.$$

Proof of Lemma 82. When $\alpha = 0$, we have for all $g \in L^r(\mathbb{T}^d)$,

$$\|g \circ \nabla f\|_{L^r(\mathbb{T}^d)}^r = \int |g(\nabla f)|^r = \int |g|^r \det(\nabla^2 f^*) \leq \lambda \|g\|_{L^r(\mathbb{T}^d)}^r.$$

To prove the claim when $\alpha = 1$, let $g \in H_0^{1,r}(\mathbb{T}^d)$, and notice that the identity $\nabla(g \circ \nabla f) = \nabla^2 f \nabla g(\nabla f)$ holds in the sense of weak derivatives. Together with Lemma 99, we deduce

$$\begin{aligned}\|g \circ \nabla f\|_{H^{1,r}(\mathbb{T}^d)}^r &\asymp \|g \circ \nabla f\|_{L^r(\mathbb{T}^d)}^r + \int \|\nabla^2 f \cdot (\nabla g(\nabla f))\|^r \\ &\lesssim_\lambda \|g\|_{L^r(\mathbb{T}^d)}^r + \int \|\nabla^2 f(\nabla f^*) \cdot \nabla g\|^r \det(\nabla^2 f^*) \lesssim_\lambda \|g\|_{H^{1,r}(\mathbb{T}^d)}^r.\end{aligned}$$

The bound for remaining values of α follows by interpolation. Indeed, notice that the operator

$$F : H^{\alpha,r}(\mathbb{T}^d) \to H^{\alpha,r}(\mathbb{T}^d), \quad Fg = g(\nabla f)$$

is well-defined and linear. Let $\|F\|_{\alpha,r} = \sup\{\|Fg\|_{H^{\alpha,r}(\mathbb{T}^d)} : g \in H^{\alpha,r}(\mathbb{T}^d)\}$ denote its operator norm. Since $H^\alpha(\mathbb{T}^d)$ is the α-complex interpolation space of $L^2(\mathbb{T}^d)$ and $H^1(\mathbb{T}^d)$ (cf. Lemma 98), we deduce (Lunardi, 2018, Theorem 2.6)

$$\|F\|_\alpha \leq \|F\|_0^{1-\alpha}\|F\|_1^\alpha \lesssim_\lambda 1.$$

The claim follows. □

The following bound can be deduced from Miranda (2013, Theorem 11,I).

Lemma 83 (Composition of Riesz Potentials). *Let $d \geq 1$. For all $\alpha_1, \alpha_2 \geq 0$, there exists a constant $C > 0$ such that for any $w_1, w_2 \in \mathbb{T}^d$,*

$$\int_{\mathbb{T}^d} \|y - w_1\|_{\mathbb{T}^d}^{\alpha_1 - d} \|y - w_2\|_{\mathbb{T}^d}^{\alpha_2 - d} dy \leq C \begin{cases} \|w_1 - w_2\|_{\mathbb{T}^d}^{\alpha_1 + \alpha_2 - d}, & \alpha_1 + \alpha_2 < d, \\ \log\big(1/\|w_1 - w_2\|\big), & \alpha_1 + \alpha_2 = d, \\ 1, & \alpha_1 + \alpha_2 > d. \end{cases}$$

We also state the following related estimate.

Lemma 84. *Let $t \in (1, d)$. Then, for any $\epsilon > 0$ and $x \in \mathbb{T}^d$,*

$$\int_{B(0,\epsilon)} \|y - x\|_{\mathbb{T}^d}^{t-d} dy \leq \epsilon^t \wedge \|x\|^{t-d} \epsilon^d.$$

Proof of Lemma 84. With the convention that $\int_\emptyset (\cdot) = 0$, we have,

$$\begin{aligned} \int_{B(0,\epsilon)} \|y - x\|_{\mathbb{T}^d}^{t-d} dy &\leq \int_{B(x,\epsilon/2)} \|y - x\|_{\mathbb{T}^d}^{t-d} dy + \int_{B(0,\epsilon) \setminus B(x,\epsilon/2)} \|y - x\|_{\mathbb{T}^d}^{t-d} dy \\ &\lesssim \int_0^{\epsilon/2} r^{t-d} r^{d-1} dr + \int_{B(0,\epsilon)} \epsilon^{t-d} dy \lesssim \epsilon^t. \end{aligned}$$

Furthermore, if $\|x\| \geq 2\epsilon$ we have

$$\begin{aligned} \int_{B(0,\epsilon)} \|y - x\|_{\mathbb{T}^d}^{t-d} dy &\leq \int_{B(0,\epsilon)} \big|\|x\| - \|y\|\big|^{t-d} dy \\ &\leq \int_0^\epsilon \big|\|x\| - r\big|^{t-d} r^{d-1} dr \lesssim \|x\|^{t-d} \int_0^\epsilon r^{d-1} dr \lesssim \|x\|^{t-d} \epsilon^d. \end{aligned}$$

This proves the claim. □

Finally, we state a multivariate version of Lyapunov's central limit theorem.

Lemma 85. *Let $\{U_{n,i} : 1 \leq i \leq n < \infty\}$ be a triangular array of independent random variables in $\mathbb{R}^d$ with mean zero and finite second moment. Define $V_n = \sum_{i=1}^n \mathrm{Cov}[U_{n,i}]$, and let $v_n^2 = \lambda_{\min}(V_n)$, for all $n = 1, 2, \dots$. If for some $\delta > 0$, Lyapunov's condition*

$$\frac{1}{v_n^{2+\delta}} \sum_{i=1}^n \mathbb{E}\Big[|U_{n,i}|^{2+\delta}\Big] = o(1)$$

holds, then

$$V_n^{-1/2} \sum_{i=1}^n U_{n,i} \xrightarrow{d} N(0, I_d).$$

Chapter 7

Efficient Inference for the Quadratic Wasserstein Distance

We now build upon the preceding two chapters to address the problem of performing *inference* for the quadratic Wasserstein distance in general dimension. As we already discussed in Chapter 1, this problem is very well-studied, and essentially all past work has aimed at finding sufficient conditions for the empirical Wasserstein distance $W_2^2(P_n, Q_m)$ to admit a limiting distribution centered at its population counterpart. However, as we know from our results in Chapter 3, the bias of the empirical Wasserstein distance typically satisfies

$$\mathbb{E}W_2^2(P_n, Q_m) - W_2^2(P, Q) \gtrsim (n \wedge m)^{-2/d}$$

for absolutely continuous probability measures P, Q in dimension d. Furthermore, its variance typically scales at the parametric rate (del Barrio and Loubes, 2019), and is thus dominated by its squared bias when $d \geq 5$. In this regime, the empirical Wasserstein distance cannot admit a non-degenerate limiting distribution centered at its population counterpart. This is a fundamental barrier for performing inference.

We take a different approach in this chapter. Rather than analyzing the empirical Wasserstein distance, we consider a family of plugin estimators based on density estimation. These estimators have already made an appearance in Chapter 5, where we derived upper bounds on their risk. Although they have several disadvantages compared to the empirical Wasserstein distance—for instance, they introduce tuning parameters, and are more challenging to compute—we will see that they are more amenable to performing inference in general dimension. Indeed, our results in Section 7.1 below show that plugin estimators based on wavelet or kernel density estimation achieve a $\sqrt{n}$-central limit theorem centered at their population counterpart, in any dimension, provided the smoothness of the underlying densities is sufficiently high. To the best of our knowledge, this leads to the first method for constructing confidence intervals for the Wasserstein distance in the regime where both measures are of dimension $d \geq 5$. We also develop the semiparametric efficiency theory for the Wasserstein distance functional, showing that our various estimators of the Wasserstein distance are asymptotically efficient in the high-smoothness regime.

One downside of our results is the fact that our density estimators require the underlying support of the measures P and Q to be known, and to be of appropriate type (specifically, rectangular or toric). We address this issue in Section 7.3, where we derive a "one-step" bias-corrected estimator of the Wasserstein distance. As we will see, this estimator allows the practitioner to plug in *any* density estimator, and makes no assumptions on the underlying domain, but continues to be semiparametric efficient under black-box rate conditions on the density estimator. We believe there is significant potential for this estimator to be adopted in practical applications.

We close this chapter with a discussion, in Section 7.4, where we pose the following question: What is the minimal smoothness condition needed to perform efficient inference for the Wasserstein distance? We will see that there exists a second-order debiased estimator with a more favourable smoothness cut-off than the plugin or one-step estimators analyzed above, but finding the optimal cut-off remains an open question.

7.1 Central Limit Theorems for Plugin Estimators

In this section, we derive limit theorems for the density plugin estimators of the Wasserstein distance which were introduced in Chapter 5. Concretely, let $P, Q \in \mathcal{P}_{\mathrm{ac}}(\Omega)$ be absolutely continuous distributions with respective densities p, q, which are either supported over the unit hypercube $\Omega = [0,1]^d$ or the d-dimensional flat torus $\Omega = \mathbb{T}^d$. Let $X_1, \dots, X_n \sim P$ and $Y_1, \dots, Y_m \sim Q$ be i.i.d. samples which are independent of each other. Recall that we respectively denote by $P_n, \widehat{P}_n^{(\mathrm{bc})}, \widehat{P}_n^{(\mathrm{ker})}$ (resp. $Q_m, \widehat{Q}_m^{(\mathrm{bc})}, \widehat{Q}_m^{(\mathrm{ker})}$), the empirical measure and the distributions induced by the boundary-corrected and kernel density estimators of p (resp. q), as defined in Sections 5.2–5.3 of Chapter 5. Given a smoothness parameter $s > 0$ to be specified, let their tuning parameters be chosen as $2^{J_n} \asymp h_n^{-1} \asymp n^{1/(d+2s)}$, and assume that the kernel K satisfies condition $\mathbf{K(2(s+1))}$ for some $\kappa > 0$. Furthermore, let φ_0 be a Brenier potential in the optimal transport problem from P to Q, and write

$$\sigma_\rho^2 = (1-\rho)\operatorname{Var}_P[\phi_0(X)] + \rho \operatorname{Var}_Q[\psi_0(Y)], \quad \text{for any } \rho \in [0,1], \tag{7.1}$$

where $\phi_0 = \|\cdot\|^2 - 2\varphi_0$ and $\psi_0 = \|\cdot\|^2 - 2\varphi_0^*$ are Kantorovich potentials induced by φ_0. Under the conditions we will make below, the Brenier potential φ_0 is uniquely-defined up to addition by a constant, thus the variances in the above display do not depend on the particular choice of φ_0.

For the various estimators $\widehat{P}_n$ and $\widehat{Q}_m$ under consideration, we will derive central limit theorems of the form

$$\sqrt{n}\Big(W_2^2(\widehat{P}_n, Q) - W_2^2(P, Q)\Big) \rightsquigarrow N(0, \sigma_0^2), \quad \text{as } n \to \infty, \quad \text{and} \tag{7.2}$$

$$\sqrt{\frac{nm}{n+m}}\Big(W_2^2(\widehat{P}_n, \widehat{Q}_m) - W_2^2(P, Q)\Big) \rightsquigarrow N(0, \sigma_\rho^2), \quad \text{as } n, m \to \infty, \ \frac{n}{n+m} \to \rho, \tag{7.3}$$

for some $\rho \in [0,1]$. Our main result is the following.

Theorem 26. Let $\Omega \in \{\mathbb{T}^d, [0,1]^d\}$. Assume that $P, Q \in \mathcal{P}_{\mathrm{ac}}(\Omega)$ admit positive and bounded densities p, q over Ω. Then, the following assertions hold.

(i) (Density Estimation over the Torus) Let $\Omega = \mathbb{T}^d$ and assume $p, q \in \mathcal{C}^s(\Omega)$ for some $s > 0$ satisfying $s > d/2 - 2$. Then, equations (7.2)–(7.3) hold when

$$(\widehat{P}_n, \widehat{Q}_m) = (\widehat{P}_n^{(\mathrm{ker})}, \widehat{Q}_m^{(\mathrm{ker})}).$$

(ii) (Density Estimation over the Hypercube) Let $\Omega = [0,1]^d$, and assume $p, q \in \mathcal{C}^s(\Omega)$ for some $s > 0$ satisfying $s > d/2 - 2$. Assume additionally that $\varphi_0 \in \mathcal{C}^{s+2}(\Omega)$. Then, equation (7.2) holds when

$$(\widehat{P}_n, \widehat{Q}_m) = (\widehat{P}_n^{(\mathrm{bc})}, \widehat{Q}_m^{(\mathrm{bc})}).$$

Furthermore, equation (7.3) holds under the additional condition $\varphi_0^* \in \mathcal{C}^{s+2}(\Omega)$.

(iii) (Empirical Measures) Let Ω be either $\mathbb{T}^d$ or $[0,1]^d$. Assume $d \leq 3$, and $\varphi_0 \in \mathcal{C}^2(\Omega)$. Then equations (7.2)–(7.3) hold when

$$(\widehat{P}_n, \widehat{Q}_m) = (P_n, Q_m).$$

To the best of our knowledge, Theorem 26 provides the first known central limit theorems for nonparametric plugin estimators of the squared Wasserstein distance in arbitrary dimension $d \geq 1$ which are centered at their population counterpart $W_2^2(P,Q)$. We emphasize that the parametric scaling in the above result is made possible by the smoothness condition $s > d/2-2$. We do not generally expect that a central limit theorem for $W_2^2(\widehat{P}_n, Q)$ centered at $W_2^2(P,Q)$ can be obtained when $s < d/2 - 2$, as the squared bias of this estimator may then dominate its variance.

We recall again that, even in the absence of smoothness conditions, del Barrio and Loubes (2019) derived limit laws of the form

$$\sqrt{n}\left(W_2^2(P_n, Q) - \mathbb{E}W_2^2(P_n, Q)\right) \rightsquigarrow N(0, \mathrm{Var}[\phi_0(X)]), \quad n \to \infty, \tag{7.4}$$

and two-sample analogues, for any $d \geq 1$. While such results are important and hold under milder regularity conditions than those of Theorem 26, their centering sequence is a barrier to their use for statistical inference for Wasserstein distances. The low-dimensional case $d \leq 3$ is an exception, in which the sequence $\mathbb{E}W_2^2(P, Q_n)$ can be replaced by $W_2^2(P,Q)$, as we show in Theorem 26(iii). This fact can be deduced from our bias bounds in Corollary 19; a similar observation for $d \leq 3$ was also made in the recent work of Hundrieser et al. (2022), under weaker assumptions than ours.

Theorem 26 is a consequence of the stability bounds in Theorem 18 and Proposition 23 in Chapter 5, which we use to show that $W_2^2(\widehat{P}_n, Q) - W_2^2(P,Q)$ asymptotically has same distribution as the linear functional $F(\widehat{P}_n) = \int \phi_0 d(\widehat{P}_n - P)$. We defer the proof to Appendix 7.A. Though our arguments differ significantly from those used by del Barrio and

Loubes (2019), this functional F also plays an important role in their work. Indeed, they prove that $n\,\mathrm{Var}[W_2^2(P_n, Q) - F(P_n)] = o(1)$ under mild conditions. In Appendix 7.C, we provide an alternate proof of Theorem 26 which does not make use of our stability bounds, but which instead combines a generalization of the proof strategy of del Barrio and Loubes (2019), together with our convergence rates for optimal transport maps in Theorems 19, 22 and 20.

The variance σ_ρ^2 is positive if and only if ϕ_0 and ψ_0 are non-constant, thus the distributional limits in Theorem 26 are non-degenerate whenever $P \neq Q$. When $P = Q$, it could already have been deduced from Lemma 106 of Chapter 5 that, for instance, the correct scaling for $W_2^2(\widehat{P}_n^{(\mathrm{bc})}, Q)$ is of larger order than $\sqrt{n}$. We leave open the question of obtaining limit laws under this regime.

The variances appearing in Theorem 26 can be consistently estimated using estimators for the Kantorovich potentials ϕ_0 and ψ_0. Indeed, using a qualitative stability result for Kantorovich potentials (Santambrogio, 2015), we show in Proposition 40 of Appendix 7.A.3 that for any of the estimators $(\widehat{P}_n, \widehat{Q}_m)$ described in Theorem 26, if $(\widehat{\phi}_{nm}, \widehat{\psi}_{nm})$ is a bounded pair of Kantorovich potentials in the optimal transport problem from $\widehat{P}_n$ to $\widehat{Q}_m$, then,

$$\widehat{\sigma}_{0,nm}^2 := \mathrm{Var}_{U\sim P_n}[\widehat{\phi}_{nm}(U)] \xrightarrow{p} \sigma_0^2, \text{ and, } \quad \widehat{\sigma}_{1,nm}^2 := \mathrm{Var}_{V\sim Q_m}[\widehat{\psi}_{nm}(V)] \xrightarrow{p} \sigma_1^2.$$

Letting $\widehat{\sigma}_{nm}^2 = \frac{m\widehat{\sigma}_{0,nm}^2 + n\widehat{\sigma}_{1,nm}^2}{n+m}$, we deduce from Theorem 26(ii) that for any $\delta \in (0,1)$,

$$W_2^2(\widehat{P}_n^{(\mathrm{bc})}, \widehat{Q}_m^{(\mathrm{bc})}) \pm \widehat{\sigma}_{nm} z_{\delta/2} \sqrt{\frac{n+m}{nm}}$$

is an asymptotic, two-sample $(1-\delta)$-confidence interval for $W_2^2(P,Q)$, assuming $P \neq Q$. Here $z_{\delta/2}$ denotes the $\delta/2$ quantile of the standard Gaussian distribution. To the best of our knowledge, this is the first practical confidence interval for the Wasserstein distance between absolutely continuous distributions in arbitrary dimension, albeit under the strong assumptions that $2(\alpha+1) > d$, and that the underlying domain Ω is known and of appropriate type.

7.2 Efficiency Lower Bounds for Estimating the Wasserstein Distance

Our aim is now to derive efficiency lower bounds, showing that the asymptotic variances in Theorem 26 cannot be improved by any other regular estimatorof $W_2^2(P,Q)$. In discussing semiparametric efficiency theory, we follow the definitions and notation of van der Vaart (1998); van der Vaart (2002). We begin with a derivation of the efficient influence function of the functional

$$\Phi_Q : \mathcal{P}(\Omega) \to \mathbb{R}, \quad \Phi_Q(P) = W_2^2(P,Q),$$

where Ω is either $\mathbb{T}^d$ or a subset of $\mathbb{R}^d$, and $Q \in \mathcal{P}_{\mathrm{ac}}(\Omega)$ is given. Santambrogio (2015, Proposition 7.17) has previously derived the first variation of this functional. The following Lemma states their result in a language suitable for our development.

Lemma 86 (Efficient Influence Function)**.** *Let Ω be $\mathbb{T}^d$ or any connected and compact subset of $\mathbb{R}^d$, and let $P, Q \in \mathcal{P}_{\mathrm{ac}}(\Omega)$. Assume that the density of at least one of P and Q is positive over Ω. Let (ϕ_0, ψ_0) denote a pair of Kantorovich potentials in the optimal transport problem from P to Q, uniquely defined up to translation by a constant, and define the map*

$$\widetilde{\Phi}_{(P,Q)}(x) = \phi_0(x) - \int \phi_0 dP, \quad x \in \Omega.$$

Let $\dot{\mathcal{P}}_P \subseteq L_0^2(P)$ be any tangent set containing $\widetilde{\Phi}_{(P,Q)}$. Then, the functional Φ_Q is differentiable relative to $\dot{\mathcal{P}}_P$, with efficient influence function given by $\widetilde{\Phi}_{(P,Q)}$.

Lemma 86 is proved in Appendix 7.B. The assumption that P or Q have support equal to Ω is only used to ensure that ϕ_0 is unique, up to translation by a constant (cf. Proposition 7.18 of Santambrogio (2015)). While this condition is not necessary (Staudt, Hundrieser, and Munk, 2022), we retain it for simplicity since we require it for our upper bounds.

By combining this result with the Convolution Theorem (van der Vaart (1998), Theorem 25.20), it immediately follows that any regular estimator sequence of $\Phi_Q(P)$ has asymptotic variance bounded below by $\mathrm{Var}_P[\phi_0(X)]/n$. The one-sample plugin estimators in Theorem 26 are thus optimal among regular estimators. A similar remark was also made in the recent independent work of Goldfeld et al. (2024), which studies efficient statistical inference for several variants of the Wasserstein distance.

We next complement this result with an asymptotic minimax lower bound, which relaxes the assumption of regularity of such estimator sequences, at the expense of only comparing their worst-case risk. In this case, we also consider the two-sample setting. Using a construction of van der Vaart (1998), we fix two differentiable paths $(P_{t,h_1})_{t\geq 0}$ and $(Q_{t,h_2})_{t\geq 0}$, for any $(h_1, h_2) \in \mathbb{R}^2$, with respective score functions $h_1\widetilde{\Phi}_{(P,Q)}$ and $h_2\widetilde{\Psi}_{(P,Q)}$, where $\widetilde{\Psi}_{(P,Q)}(y) := \psi_0(y) - \int \psi_0 dQ$. These paths are defined in equations (7.18–7.19) of Appendix 7.B, and we use them to obtain the following asymptotic minimax lower bound.

Theorem 27 (Asymptotic Minimax Lower Bound over $\mathbb{T}^d$)**.** Given $M, \gamma > 0$ and $\alpha > 1$, let $P, Q \in \mathcal{P}_{\mathrm{ac}}(\mathbb{T}^d)$ admit densities $p, q \in \mathcal{C}^{\alpha-1}(\mathbb{T}^d; M, \gamma)$. Let (ϕ_0, ψ_0) denote a pair of Kantorovich potentials between P and Q, unique up to translation by a constant. Then, there exist $\overline{M}, \overline{\gamma}, \overline{u} > 0$, depending only on M, γ, α, such that P_{t,h_1} and Q_{t,h_2} admit densities in $\mathcal{C}^{\alpha-1}(\mathbb{T}^d; \overline{M}, \overline{\gamma})$, for all $t > 0$ and $h_1, h_2 \in \mathbb{R}$ satisfying $t(|h_1| \vee |h_2|) \leq \overline{u}$. Furthermore,

(i) (One Sample) For any estimator sequence $(U_n)_{n\geq 1}$, we have

$$\sup_{\substack{\mathcal{I}\subseteq\mathbb{R} \\ |\mathcal{I}|<\infty}} \liminf_{n\to\infty} \sup_{h\in\mathcal{I}} n\mathbb{E}_{n,h}\big|U_n - \Phi_Q(P_{n^{-1/2},h})\big|^2 \geq \mathrm{Var}_P[\phi_0(X)].$$

where $\mathbb{E}_{n,h}$ denotes the expectation taken over the probability measure $P_{n^{-1/2},h}^{\otimes n}$.

(ii) (Two Sample) For any estimator sequence $(U_{nm})_{n,m\geq 1}$, we have

$$\sup_{\substack{\mathcal{I}\subseteq\mathbb{R}^2\\|\mathcal{I}|<\infty}} \liminf_{n,m\to\infty} \sup_{(h_1,h_2)\in\mathcal{I}} \frac{nm}{n+m}\mathbb{E}_{n,m,h_1,h_2}\left|U_{nm} - W_2^2(P_{n^{-1/2},h_1}, Q_{m^{-1/2},h_2})\right|^2 \\ \geq (1-\rho)\operatorname{Var}_P[\phi_0(X)] + \rho\operatorname{Var}_Q[\psi_0(Y)],$$

where the limit inferior is taken as $n/(n+m)\to\rho\in[0,1]$, and $\mathbb{E}_{n,m,h_1,h_2}$ denotes the expectation taken over the probability measure $P^{\otimes n}_{n^{-1/2},h_1}\otimes Q^{\otimes m}_{m^{-1/2},h_2}$.

The proof of Theorem 27 appears in Appendix 7.B. For technical purposes, our statement assumes that P,Q admit densities lying in a strict subset $\mathcal{C}^{\alpha-1}(\mathbb{T}^d;M,\gamma)$ of $\mathcal{C}^{\alpha-1}(\mathbb{T}^d;\overline{M},\overline{\gamma})$, the latter being the class in which our differentiable paths are shown to lie. With this caveat, our plugin estimators achieve the asymptotic minimax lower bounds of Theorem 27. For example, under the conditions of Theorem 20, when $2(\alpha+1)>d$ we deduce that

$$\sup_{\substack{\mathcal{I}\subseteq\mathbb{R}^2\\|\mathcal{I}|<\infty}} \liminf_{n,m\to\infty} \sup_{(h_1,h_2)\in\mathcal{I}} \frac{nm}{n+m}\mathbb{E}_{n,m,h_1,h_2}\left|W_2^2(\widehat{P}_n^{(\mathrm{ker})},\widehat{Q}_m^{(\mathrm{ker})}) - W_2^2(P_{n^{-1/2},h_1}, Q_{m^{-1/2},h_2})\right|^2 \\ = (1-\rho)\operatorname{Var}_P[\phi_0(X)] + \rho\operatorname{Var}_Q[\psi_0(Y)].$$

It can be verified that Theorem 27 continues to hold with $\mathbb{T}^d$ replaced by $[0,1]^d$, under the additional condition that $\varphi_0,\varphi_0^*\in\mathcal{C}^{\alpha+1}([0,1]^d)$. We were unable, however, to derive differentiable paths $Q_{t,h}=(\nabla\varphi_{t,h})_{\#}P_{t,h}$ which simultaneously satisfy the Hölder continuity properties of Theorem 27 while also having Brenier potentials $\varphi_{t,h},\varphi_{t,h}^*$ with uniformly bounded $\mathcal{C}^{\alpha+1}([0,1]^d)$ norm.

7.3 A One-Step Estimator

The previous two sections established that several plugin estimators of the Wasserstein distance are semiparametric efficient when the smoothness of the underlying densities is sufficiently high. A possibly surprising feature of these results was the fact that *undersmoothing was not necessary*: the bandwidth used in our density estimators was always taken to be the minimax-optimal choice. This forced us to make strong structural assumptions which ensured that our plugin estimators did not have leading-order bias for all $s>0$.

In this section, we aim to relax some of these assumptions by instead proposing a "one-step" debiased estimator of the Wasserstein functional. First-order estimators of this type have been at the heart of many recent developments in the causal inference literature, and more broadly in a variety of semiparametric inference problems (see for instance Kennedy (2022) for a survey). In such problems, one-step estimators are typically favored over plugin estimators because they allow for nuisance parameters of a functional to be estimated in a black-box fashion, while still retaining asymptotic efficiency. We will see that this general story applies to the Wasserstein functional: treating the underlying densities and Kantorovich potentials as nonparametric nuisance parameters, and given estimators of these quantities, we will construct a one-step estimator which is agnostic to the structure of the nuisance estimates, and is semiparametric efficient under black-box rate conditions.

Let us briefly explain why our estimator involves the estimation of Kantorovich potentials in addition to the densities. Recall that in Proposition 23, we derived a two-sample von Mises expansion of the W_2^2 functional. Given preliminary density estimator $\widehat{P}_n$ and $\widehat{Q}_m$, this expansion roughly suggests that

$$W_2^2(\widehat{P}_n, \widehat{Q}_m) \approx W_2^2(P, Q) + \int \phi_0 d(\widehat{P}_n - P) + \int \psi_0 d(\widehat{Q}_m - Q),$$

where the approximation holds up to an error that decays quadratically with respect to the Wasserstein risk of $\widehat{P}_n$ and $\widehat{Q}_m$. The above display suggests that the leading-order bias of $W_2^2(\widehat{P}_n, \widehat{Q}_m)$ is on the order of

$$\mathbb{E}\left[\int \phi_0 d(\widehat{P}_n - P) + \int \psi_0 d(\widehat{Q}_m - Q)\right].$$

Although we were able to show that this bias term is negligible for the plugin estimators studied in Section 7.1, our results required strong structural conditions on the density estimators and underlying support. There is no a priori reason to expect that the above bias term is negligible for general density estimators, and we will instead construct an estimator which subtracts these bias terms. This will require us to estimate the underlying nuisance functions ϕ_0 and ψ_0.

Let us now provide a formal construction of our estimator. Let $\Omega \subseteq \mathbb{R}^d$ be a closed, convex set. We continue to denote by P, Q two absolutely continuous probability distributions, with respective densities p, q over Ω. Assume the practitioner has access to two independent i.i.d. samples

$$X_1, \dots, X_{2n} \sim P, \quad Y_1, \dots, Y_{2m} \sim Q,$$

where we assume the sample sizes $2n$ and $2m$ are even for simplicity. We split the samples into the sets

$$\begin{aligned} &\mathbf{X}_{(1)} = \{X_1, \dots, X_n\}, \quad \mathbf{X}_{(2)} = \{X_{n+1}, \dots, X_{2n}\}, \text{ and} \\ &\mathbf{Y}_{(1)} = \{Y_1, \dots, Y_n\}, \quad \mathbf{Y}_{(2)} = \{Y_{n+1}, \dots, Y_{2n}\}, \end{aligned}$$

and we assume that the following nuisance function estimators are given:

- $\widehat{p}_n$ is an estimator of the density p based on $\mathbf{X}_{(1)}$. We assume that $\widehat{p}_n$ is a proper density over Ω, and we let $\widehat{P}_n$ be the probability distribution with density $\widehat{p}_n$;
- $\widehat{q}_m$ and $\widehat{Q}_m$ are defined similarly, based on the sample $\mathbf{Y}_{(1)}$;
- $\widehat{T}_n = \nabla\widehat{\varphi}_n$ is an estimator of the Brenier map T_0 based on $\mathbf{X}_{(1)}$, which is proper in the sense that it is the gradient of a convex function $\widehat{\varphi}_n$. We assume this convex function is additively normalized such that the induced Kantorovich potential

$$\widehat{\phi}_n = \|\cdot\|^2 - 2\widehat{\varphi}_n$$

satisfies $\int_\Omega \widehat{\phi}_n d\widehat{P}_n = 0$.

- $\widehat{S}_m = \nabla \widehat{\eta}_m$ is an estimator of the inverse map T_0^{-1} based on $\mathbf{Y}_{(1)}$, where again we assume $\widehat{\eta}_m$ is a convex function such that the induced Kantorovich potential
$$\widehat{\psi}_m = \| \cdot \|^2 - 2\widehat{\eta}_m$$
satisfies $\int_\Omega \widehat{\psi}_m d\widehat{Q}_m = 0$.

Furthermore, define the following empirical measures based on $\mathbf{X}_{(2)}$ and $\mathbf{Y}_{(2)}$,

$$P_n' = \frac{1}{n} \sum_{i=n+1}^{2n} \delta_{X_i}, \quad Q_m' = \frac{1}{m} \sum_{j=m+1}^{2m} \delta_{Y_j}.$$

We then arrive at the following one- and two-sample estimators of the quadratic optimal transport cost

$$\widehat{W}_n = W_2^2(\widehat{P}_n, Q) + \int \widehat{\phi}_n dP_n', \text{ and}$$
$$\widehat{W}_{nm} = W_2^2(\widehat{P}_n, \widehat{Q}_m) + \int \widehat{\phi}_n dP_n' + \int \widehat{\psi}_m dQ_m'.$$

We are now ready to the state the main result of this section.

Theorem 28. *Let $\Omega \subseteq \mathbb{R}^d$ be a closed, convex set. Let $P, Q \in \mathcal{P}_{\mathrm{ac}}(\Omega) \cap \mathcal{P}_2(\Omega)$ be absolutely continuous distributions admitting respective densities $p, q \in L^\infty(\Omega)$, and let T_0 be the Brenier map pushing P forward onto Q. Let σ_ρ^2 be defined as in equation* (7.1).

(i) (One Sample) Assume that T_0 is Lipschitz over Ω, and that there exists a pair of Kantorovich potentials (ϕ_0, ψ_0) in the optimal transport problem from P to Q such that

$$\|\widehat{p}_n\|_{L^\infty(\Omega)} = O_p(1), \quad \inf_{a \in \mathbb{R}} \|\widehat{\phi}_n - \phi_0 - a\|_{L^2(\Omega)} = o_p(1), \text{ and} \tag{7.5}$$

$$W_2^2(\widehat{P}_n, P) + W_2(\widehat{P}_n, P)\|\widehat{T}_n - T_0\|_{L^2(\Omega)} = o_p(n^{-1/2}). \tag{7.6}$$

Then, as $n \to \infty$,
$$\sqrt{n}\big(\widehat{W}_n - W_2^2(P, Q)\big) \xrightarrow{d} N(0, \sigma_0^2).$$

(ii) (Two Sample) Assume that T_0 is bi-Lipschitz over Ω. Assume further that there exists a pair of Kantorovich potentials (ϕ_0, ψ_0) in the optimal transport problem from P to Q such that assumptions (7.5–7.6) hold, and

$$\|\widehat{q}_m\|_{L^\infty(\Omega)} = O_p(1), \quad \inf_{a \in \mathbb{R}} \|\widehat{\psi}_m - \psi_0 - a\|_{L^2(\Omega)} = o_p(1), \text{ and}$$
$$W_2^2(\widehat{Q}_m, Q) + W_2(\widehat{Q}_m, Q)\|\widehat{S}_m - T_0^{-1}\|_{L^2(\Omega)} = o_p(m^{-1/2}).$$

Then, as $n, m \to \infty$ such that $n/m \to \rho \in (0, 1)$, it holds that

$$\sqrt{\frac{nm}{n+m}}\big(\widehat{W}_{nm} - W_2^2(P, Q)\big) \xrightarrow{d} N(0, \sigma_\rho^2).$$

We make several remarks regarding Theorem 28.

- The estimator $\widehat{W}_n$ is constructed with a sample of size $2n$, but achieves the same limiting variance as our plugin estimators based on a sample of size n. Therefore, $\widehat{W}_n$ itself is inefficient, but there is a simple modification of this estimator which is efficient. Let $\widehat{W}_n'$ be the estimator obtained through the same procedure as $\widehat{W}_n$, but with the nuisance functions estimated on the second half of the data $\mathbf{X}_{(2)}$, and with P_n' replaced by the empirical measure on the first half of the data $\mathbf{X}_{(1)}$. Then, the estimator
$$\frac{1}{2}\widehat{W}_n + \frac{1}{2}\widehat{W}_n'$$
can be shown to be efficient under similar conditions as those of Theorem 28 (a similar procedure is used by Bickel and Ritov (1988)). The two-sample case can also be handled similarly, but we omit formal statements in order to alleviate notation.
- The most substantive assumption in Theorem 28 is the rate condition (7.6). This assumption is reminiscent of that of doubly robust estimators in the causal inference literature (Chernozhukov et al., 2018), which would typically require that the mere product of errors to decay at a fast rate,
$$W_2(\widehat{P}_n, P)\|\widehat{T}_n - T_0\|_{L^2(\Omega)} = o_p(n^{-1/2}).$$
In our setting, we additionally require the condition $W_2(\widehat{P}_n, P) = o_p(n^{-1/4})$. Roughly speaking, this means that the underlying densities must be estimated well (at least at the $n^{-1/4}$ rate), but that the nuisance estimator $\widehat{\phi}_n$ does not need to be estimated at a fast rate. For example, if the density P is estimable at the parametric rate in Wasserstein distance, then it suffices to choose $\widehat{\phi}_n$ such that the fitted map $\widehat{T}_n$ is consistent in $L^2(\Omega)$. This is a very weak requirement; for instance, Segers (2022) derives uniformly consistent estimators of optimal transport maps under almost no regularity assumptions.
- Notice carefully that Theorem 28 does not make the assumption that the underlying Kantorovich potentials (ϕ_0, ψ_0) are unique. Instead, we merely assume that there exists a pair of nuisance estimates $(\widehat{\phi}_n, \widehat{\psi}_m)$ which consistently estimates some allowable pair of population Kantorovich potentials (ϕ_0, ψ_0), up to additive normalization, and the limiting variances σ_ρ^2 are to be understood with respect to this pair.

Let us now turn to the proof of Theorem 28.

7.3.1 Proof of Theorem 28

For the purpose of this proof only, let us make use of the following characterization of the Sobolev seminorms: for any maps $f \in H^1(\Omega)$ and $g \in L_0^2(\Omega)$,
$$\|f\|_{\dot{H}^1(\Omega)} = \int_\Omega \|\nabla f\| d\mathcal{L}, \text{ and,}$$
$$\|g\|_{\dot{H}^{-1}(\Omega)} = \sup\left\{\int_\Omega gf : \|f\|_{H^1(\Omega)} \leq 1, f \in H^1(\Omega)\right\}.$$

We will also make use of the following result due to Loeper (2006); Peyre (2018).

Lemma 87. *Assume that $\mu, \nu \in \mathcal{P}_{\mathrm{ac}}(\Omega) \cap \mathcal{P}_2(\Omega)$ admit densities $f, g \in L^\infty(\Omega)$. Then, there is a constant $C = C(\|f\|_{L^\infty(\Omega)}, \|g\|_{L^\infty(\Omega)}, \Omega, d)$ such that for any map $h \in H^1(\Omega)$,*

$$\int h d(\mu - \nu) \leq C \|h\|_{\dot{H}^1(\Omega)} W_2(\mu, \nu).$$

Let us now turn to the proof. We will prove the first claim, and a similar argument may be used to deduce the second. We have,

$$\begin{aligned}
&\widehat{W}_n^{(1)} - W_2^2(P,Q) \\
&\quad = W_2^2(\widehat{P}_n, Q) - W_2^2(P,Q) + \int \widehat{\phi}_n dP_n' \\
&\quad = \int \phi_0 d(\widehat{P}_n - P) + \int \widehat{\phi}_n dP_n' + o_p(n^{-1/2}),
\end{aligned}$$

where we used the fact that, by Theorem 18 of Chapter 5 (and remarks thereafter), the order assessment

$$W_2^2(\widehat{P}_n, Q) - W_2^2(P,Q) - \int \phi_0 d(\widehat{P}_n - P) = O_p\big(W_2^2(\widehat{P}_n, P)\big) = o_p(n^{-1/2})$$

holds under the Lipschitz assumption on T_0, and under our assumptions on the nuisance parameter estimators. Furthermore, one has,

$$\begin{aligned}
&\int \phi_0 d(\widehat{P}_n - P) + \int \widehat{\phi}_n dP_n' \\
&\quad = \int (\phi_0 - \widehat{\phi}_n) d(\widehat{P}_n - P) + \int \widehat{\phi}_n d(P_n' - P).
\end{aligned}$$

where we used the fact that $\int \widehat{\phi}_n d\widehat{P}_n = 0$. Notice that

$$\int (\phi_0 - \widehat{\phi}_n) d(\widehat{P}_n - P) \leq \|\widehat{\phi}_n - \phi_0\|_{\dot{H}^1(\Omega)} \|\widehat{p}_n - p\|_{\dot{H}^{-1}(\Omega)} \lesssim \|\widehat{T}_n - T_0\|_{L^2(\Omega)} W_2(\widehat{P}_n, P),$$

where we used Lemma 87, together with the fact that $\widehat{p}_n$ (and p) is bounded over Ω with probability tending to one. By assumption on the nuisance parameter estimators, we thus have

$$\int (\phi_0 - \widehat{\phi}_n) d(\widehat{P}_n - P) = o_p(n^{-1/2}).$$

Combining these facts, we arrive at

$$\begin{aligned}
\sqrt{n}\big(\widehat{W}_n - W_2^2(P,Q)\big) &= \sqrt{n} \int \widehat{\phi}_n d(P_n' - P) + o_p(1) \\
&= \sqrt{n} \int \phi_0 d(P_n' - P) + \sqrt{n} \int (\widehat{\phi}_n - \phi_0) d(P_n' - P) + o_p(n^{-1/2}).
\end{aligned}$$

The first term in the final line of the above display converges in distribution to $N(0, \sigma_0^2)$ by the classical central limit theorem. It thus remains to show that the second term vanishes in probability. To this end, notice that by independence of P_n' and $\widehat{\phi}_n$, it holds (cf. Lemma 2 of Kennedy, Balakrishnan, and G'Sell (2020))

$$\sqrt{n}\int(\widehat{\phi}_n - \phi_0)d(P_n' - P) = O_p\left(\sqrt{\mathrm{Var}[\widehat{\phi}_n - \phi_0 | X_1, \dots, X_n]}\right).$$

Using the fact that P has a bounded density over Ω, we have

$$\mathrm{Var}[\widehat{\phi}_n - \phi_0 | X_1, \dots, X_n] \leq \inf_{a\in\mathbb{R}} \|\widehat{\phi}_n - \phi_0 - a\|_{L^2(\Omega)}^2 = o_p(1),$$

by assumption. The claim now follows. □

7.4 Toward Higher-Order Estimation

The one-step estimator presented in the previous section allows for efficient inference to be performed over a wide range of model classes, including but not limited to the Hölder models used in Section 7.1. Nevertheless, if we restrict our attention to these Hölder models, it is natural to ask whether the plugin or one-step estimators can be improved by higher-order debiasing. Indeed, it has well-known that for other traditional functional estimation problems, plugin and one-step estimators are minimax suboptimal, and can be improved by debiasing the one-step estimator (Bickel and Ritov, 1988; Birgé and Massart, 1995; Laurent, 1996; Kerkyacharian and Picard, 1996). In this section, we take a first step suggesting that the same conclusion is true of the Wasserstein distance: we will derive a second-order debiased estimator which achieves a polynomially faster convergence rate than that of our plugin or one-step estimators.

For simplicity, we will focus on the simplest setting of periodic measures. Thus, we assume throughout the remainder of this section that P and Q are absolutely continuous measures over $\mathbb{T}^d$, with respective densities p, q which are bounded from below over $\mathbb{T}^d$ by a positive constant. In this case, there exists a unique pair of Kantorovich potentials (ϕ_0, ψ_0) (up to additive normalization) in the optimal transport problem from P to Q, and we fix such a pair with the normalization constant chosen such that $\int \phi_0 dP = 0$. Furthermore, for simplicity of exposition, we focus only on the one-sample case where Q is known, and an i.i.d. sample $X_1, \dots, X_{2n} \sim P$ is given.

Our main technical contribution will be to derive a second-order von Mises expansion of the Wasserstein functional, and to show that the remainder of this expansion decays cubically. We have already shown that the first-order influence function of W_2^2 is the Kantorovich potential ϕ_0. This suggests that its second-order influence function, if it were to exist, could be obtained by taking the influence function of the mapping $p \mapsto \phi_0(x)$, at any fixed point $x \in \mathbb{T}^d$. Conveniently, this mapping was our main object of study in Chapter 6, where we used a linearization of the Monge-Ampère equation to quantify its first variation. As a reminder, let us recall a result that can be deduced from Section 6.3 (and recall notation (1.18)).

Lemma 88. *Let $p, \widehat{p}, q \in \mathcal{C}_+^{2+\beta}(\mathbb{T}^d)$ for some $\beta > 2$, and let $\widehat{\varphi}$ be the unique Brenier potential from $\widehat{p}$ to q which has mean zero over the unit hypercube. Let $\widehat{\phi} = \|\cdot\|^2 - 2\varphi_0$. Furthermore, let E be operator defined in Lemma 58, namely*

$$E : H_0^2(\mathbb{T}^d) \to L_0^2(\mathbb{T}^d), \quad Eu = -\mathrm{div}(A\nabla u),$$

where $A = p \cdot (\nabla^2\varphi_0)^{-1}$. Then, there is a constant $C = C(\omega_{2+\beta}(p, \widehat{p}, q), \beta, d) > 0$ such that

$$\left\| \widehat{\phi} - \phi_0 - E^{-1}[\widehat{p} - p] \right\| \lesssim \|\widehat{p} - p\|^2_{\mathcal{C}^{1+\beta}(\mathbb{T}^d)}.$$

Using this observation, we arrive at the following result. In what follows, we recall from Proposition 39 that the operator E admits a unique periodic Green's function $\Gamma \in L^1(\mathbb{T}^d \times \mathbb{T}^d)$, which is symmetric and integrates to zero in each of its coordinates. It satisfies the identity

$$E^{-1}f(x) = \int_{\mathbb{T}^d} \Gamma(y, x) f(y) dy, \quad \text{for all } f \in \mathcal{C}_0^\infty(\mathbb{T}^d),\ x \in \mathbb{T}^d.$$

Theorem 29. *Assume the same conditions as in Lemma 88. Then, there is a constant $C > 0$ depending only on $\omega_{2+\beta}(p, \widehat{p}, q), d, \beta$ such that*

$$\begin{aligned} \Bigg| W_2^2(\widehat{P}, Q) - W_2^2(P, Q) - \int \phi_0 d(\widehat{P} - P) - \iint \Gamma d(\widehat{P} - P) d(\widehat{P} - P) \Bigg| \\ \leq C \|\widehat{p} - p\|_{\mathcal{C}^\beta(\mathbb{T}^d)} \|\widehat{p} - p\|^2_{H^{-1}(\mathbb{T}^d)}. \end{aligned}$$

This result marks an important departure between the Wasserstein distance and other familiar functionals, for which lack of existence of second-order influence functions is typical (Robins et al., 2009). For such functionals, it is usually possible to use the idea of a quadratic expansion to debias one-step estimators by first restricting the functional to a finite-dimensional subspace, where a second-order influence function may exist, and then increasing the dimension of this subspace, incurring a bias-variance trade-off. For us, the situation will be simpler, since Theorem 29 shows that the quadratic term in the von Mises expansion of $W_2^2(\widehat{P}, Q)$ is already representable by a bilinear form, with "second-order influence function" $\Gamma \in L^1(\mathbb{T}^d \times \mathbb{T}^d)$. We will thus be able to perform second-order debiasing by simply estimating Γ and the corresponding bilinear form. Nevertheless, let us emphasize that Γ does *not* lie in $L^2(\mathbb{T}^d \times \mathbb{T}^d)$, and thus is not strictly-speaking a second-order influence function of $W_2^2(\cdot, Q)$ in the sense of Robins et al. (2009).

Guided by Theorem 29, let us now define our second-order debiased estimator. We will assume as in the previous section that the i.i.d. sample is split into the sets

$$\mathbf{X}_{(1)} = \{X_1, \dots, X_n\}, \quad \mathbf{X}_{(2)} = \{X_{n+1}, \dots, X_{2n}\}.$$

Let

$$\widehat{p}_n(x) = \int_{\mathbb{R}^d} K_{h_n}(x - y) dP_n(y), \quad x \in \mathbb{T}^d$$

be a kernel density estimator based on the empirical measure $P_n = (1/n)\sum_{i=1}^n \delta_{X_i}$ from $\mathbf{X}_{(1)}$. The above definition is to be understood with the same conventions as in Chapters 5–6; in particular, $K \in \mathcal{C}_c^\infty(\mathbb{R}^d)$ is a kernel, and $K_{h_n} = K(\cdot/h_n)/h_n^d$ for some bandwidth $h_n > 0$. Furthermore, let $\widehat{\varphi}_n$ be the unique Brenier potential in the optimal transport problem from $\widehat{P}_n$ to Q, admitting mean zero over the unit hypercube:

$$\int_{[0,1]^d} \widehat{\varphi}_n(x)dx = 0.$$

Furthermore, let $\widehat{\phi}_n = \|\cdot\|^2 - 2\widehat{\varphi}_n$ be an induced Kantorovich potential. Now, define the operator

$$\widehat{E}_n : H_0^2(\mathbb{T}^d) \to L_0^2(\mathbb{T}^d), \quad \widehat{E}_n u = -\mathrm{div}(\widehat{A}_n \nabla u),$$

with $\widehat{A}_n = \widehat{p}_n \cdot (\nabla^2 \widehat{\varphi}_n)^{-1}$. Here and throughout what follows, all of our definitions are tacitly made over the event that $\widehat{p}_n$ defines a probability density on $\mathbb{T}^d$ which is bounded away from zero. Reasoning as in Chapter 6, the above operator admits a discrete and real spectrum; we denote by $0 < \lambda_1 \leq \lambda_2 \leq \dots$ its ordered sequence of eigenvalues, with corresponding eigenfunctions $\{\eta_j\}_{j=1}^\infty$ chosen to form an orthonormal basis of $L_0^2(\mathbb{T}^d)$, and such that $\widehat{E}_n \eta_j = \lambda_j \eta_j$ for all $j \geq 1$. We emphasize that these eigenfunctions and eigenvalues are random and depend on n, but we omit this dependence from our notation for ease of exposition.

Let $\widehat{\Gamma}_n$ be the periodic Green's function associated with the above operator, and for some nonnegative sequence $(J_n)_{n\geq 1}$, define the estimator

$$\widetilde{\Gamma}_n(x,y) = \sum_{j=1}^{J_n} \frac{\eta_j(x)\eta_j(y)}{\lambda_j}, \quad x, y \in \mathbb{T}^d.$$

Now, let $P_n' = (1/n)\sum_{i=n+1}^{2n} \delta_{X_i}$ be the empirical measure based on $\mathbf{X}_{(2)}$, and define the U-statistic operator based on $\mathbf{X}_{(2)}$,

$$\mathcal{U}_n g = \frac{1}{\binom{n}{2}} \sum_{n+1 \leq i < j \leq 2n} g(X_i, X_j),$$

where $g : \mathbb{T}^d \times \mathbb{T}^d \to \mathbb{R}$ is any Borel-measurable map. Our second-order debiased estimator is then defined by

$$\overline{W}_n := W_2^2(\widehat{P}_n, Q) + \int_{\mathbb{T}^d} \widehat{\phi}_n d(P_n' - \widehat{P}_n) + \mathcal{U}_n g_n,$$

where $g_n : \mathbb{T}^d \times \mathbb{T}^d \to \mathbb{R}$ is the kernel defined by

$$g_n(x,y) = \widetilde{\Gamma}_n(y,x) - \iint \Big[\widetilde{\Gamma}_n(y,x') + \widetilde{\Gamma}_n(y',x) - \widetilde{\Gamma}_n(y',x')\Big] d\widehat{P}_n(x') d\widehat{P}_n(y'), \quad x, y \in \mathbb{T}^d.$$

Our main result is the following.

Theorem 30. *Let $p, q \in \mathcal{C}^s_+(\mathbb{T}^d)$, $d \geq 3$, and assume*

$$s > \max\big\{d/4 - 1, 2\big\}. \tag{7.7}$$

Assume further that the kernel K satisfies condition $\mathbf{K(s+1)}$. *Let*

$$h_n \asymp n^{-\frac{1}{2s+d}}, \quad \textit{and} \quad J_n^{1/d} \asymp n^{\frac{2}{4s+d}}.$$

Then, it holds that

$$\sqrt{n}\big(\overline{W}_n - W_2^2(P,Q)\big) \xrightarrow{d} N(0, \sigma_0^2), \quad \textit{as } n \to \infty,$$

where σ_0^2 is defined as in equation (7.1).

Theorem 30 shows that second-order debiasing leads to semiparametric efficient inference under the smoothness condition (7.7), which is significantly weaker than that of our plugin estimators in Section 7.1 when $s > 2$. It is easy to see that this condition is also strictly weaker than what is achievable by the one-step estimator in Section 7.3. We leave it as an exciting open question to determine whether this condition is sharp, or whether it can be improved by a more careful analysis or higher-order debiasing.

Let us now turn to proving Theorems 29–30 in turn.

7.4.1 Proof of Theorem 29

Let $\epsilon \in [0,1]$, $P_\epsilon = P + \epsilon(\widehat{P} - P)$, and let φ_ϵ be the unique Brenier potential admitting mean zero over $[0,1]^d$, whose gradient pushes forward P_ϵ onto Q. Let ϕ_ϵ be the induced Kantorovich potential. Furthermore, let E_ϵ be the operator defined by

$$E_\epsilon : H_0^2(\mathbb{T}^d) \to L_0^2(\mathbb{T}^d), \quad E_\epsilon u = -\mathrm{div}(A_\epsilon \nabla u),$$

where $A_\epsilon = p_\epsilon \cdot (\nabla^2 \varphi_\epsilon)^{-1}$. Notice that, under the conditions we have placed on $p, \widehat{p}, q$, the operator E_ϵ satisfies the properties set forth in Section 6.A of Chapter 6. Define

$$g : [0,1] \to \mathbb{R}, \quad g(\epsilon) = W_2^2(P_\epsilon, Q).$$

We have already established that

$$g'(\epsilon) = \lim_{h \to 0} \frac{W_2^2(P_{\epsilon+h}, Q) - W_2^2(P_\epsilon, Q)}{h} = \int \phi_\epsilon d(\widehat{P} - P).$$

In order to compute the second-order derivative of g, write

$$\begin{aligned} &\frac{1}{h}\int (\phi_{\epsilon+h} - \phi_\epsilon) d(\widehat{P} - P) \qquad (7.8)\\ &= \frac{1}{h}\int E_\epsilon^{-1}[p_{\epsilon+h} - p_\epsilon] d(\widehat{P} - P) + \frac{1}{h}\int \Big(\phi_{\epsilon+h} - \phi_\epsilon - E_\epsilon^{-1}[p_{\epsilon+h} - p_\epsilon]\Big) d(\widehat{P} - P). \end{aligned}$$

Now, since $p_\epsilon \in \mathcal{C}_+^{2+\beta}(\mathbb{T}^d)$, with Hölder norm and density lower bound which are uniform in ϵ, we may apply Lemma 88 to the densities $p_\epsilon, p_{\epsilon+h}, q$ to deduce

$$\left\|\phi_{\epsilon+h} - \phi_\epsilon - E_\epsilon^{-1}[p_{\epsilon+h} - p_\epsilon]\right\|_{\mathcal{C}^{2+\beta}(\mathbb{T}^d)} \lesssim \|p_{\epsilon+h} - p_\epsilon\|_{\mathcal{C}^{1+\beta}(\mathbb{T}^d)}^2.$$

Combining this inequality with a simple duality argument, we obtain

$$\begin{aligned}
&\left|\int \Big(\phi_{\epsilon+h} - \phi_\epsilon - E_\epsilon^{-1}[p_{\epsilon+h} - p_\epsilon]\Big) d(\widehat{P} - P)\right| \\
&\quad\lesssim \left\|\phi_{\epsilon+h} - \phi_\epsilon - E_\epsilon^{-1}[p_{\epsilon+h} - p_\epsilon]\right\|_{H^{2+\beta}(\mathbb{T}^d)} \|\widehat{p} - p\|_{H^{-(2+\beta)}(\mathbb{T}^d)} \\
&\quad\lesssim \|p_{\epsilon+h} - p_\epsilon\|_{\mathcal{C}^{1+\beta}(\mathbb{T}^d)}^2 \|\widehat{p} - p\|_{H^{-(2+\beta)}(\mathbb{T}^d)} \lesssim h^2.
\end{aligned}$$

Returning to equation (7.8), we deduce

$$\begin{aligned}
g''(\epsilon) &= \lim_{h\to 0} \frac{1}{h} \int (\phi_{\epsilon+h} - \phi_\epsilon) d(\widehat{P} - P) \\
&= \lim_{h\to 0} \frac{1}{h} \int E_\epsilon^{-1}[p_{\epsilon+h} - p_\epsilon] d(\widehat{P} - P) \\
&= \int E_\epsilon^{-1}[\widehat{p} - p] d(\widehat{P} - P),
\end{aligned}$$

where we used the fact that $p_{\epsilon+h} - p_\epsilon = h(\widehat{p} - p)$. We have thus shown that g is twice differentiable. One may equivalently express the second derivative in terms of the periodic Green's function Γ_ϵ associated to the operator E_ϵ,

$$g''(\epsilon) = \int\int \Gamma_\epsilon(x, y) d(\widehat{P} - P)(y) d(\widehat{P} - P)(x).$$

It now follows from a second-order Taylor expansion that

$$|g(1) - g(0) - g'(0) - g''(0)| \leq \sup_{\epsilon,\epsilon' \in [0,1]} |g''(\epsilon) - g''(\epsilon')|.$$

To prove the claim, it thus remains to bound the quantity on the right-hand of the above display. Notice that for any given $\epsilon, \epsilon' \in [0, 1]$,

$$g''(\epsilon) - g''(\epsilon') = \int (u - v) d(\widehat{P} - P),$$

where $u = E_\epsilon^{-1}[\widehat{p} - p]$ and $v = E_{\epsilon'}^{-1}[\widehat{p} - p]$. We have

$$E_\epsilon u = \widehat{p} - p = E_{\epsilon'} v,$$

whence $E_{\epsilon'}[v - u] = (E_\epsilon - E_{\epsilon'})u$, and so, for any $x \in \mathbb{T}^d$,

$$v(x) - u(x) = E_{\epsilon'}^{-1}\big[(E_\epsilon - E_{\epsilon'})u\big](x)$$

$$\begin{aligned}
&= -\int \Gamma_{\epsilon'}(\cdot,x)\Big(\mathrm{div}(A_\epsilon \nabla u) - \mathrm{div}(A_{\epsilon'}\nabla u)\Big) \\
&= \int \nabla_y \Gamma_{\epsilon'}(y,x)^\top \big(A_\epsilon - A_{\epsilon'}\big)(y)\nabla u(y)dy.
\end{aligned}$$

We thus have,

$$\begin{aligned}
g''(\epsilon) - g''(\epsilon) &= \int\int \nabla_y \Gamma_{\epsilon'}(y,x)^\top \big(A_\epsilon - A_{\epsilon'}\big)(y)\nabla u(y)dydx \\
&= \int \nabla E_{\epsilon'}^{-1}[\widehat{p}-p](y)\big(A_\epsilon - A_{\epsilon'}\big)(y)\nabla u(y)dy,
\end{aligned}$$

where the final equality follows from Fubini's theorem. Deduce that for any fixed $\beta > 0$,

$$\begin{aligned}
|g''(\epsilon) - g''(\epsilon')| &\lesssim \big\|\nabla E_\epsilon^{-1}[\widehat{p}-p]\big\|_{L^2(\mathbb{T}^d)} \|A_\epsilon - A_{\epsilon'}\|_{L^\infty(\mathbb{T}^d)} \big\|\nabla u\big\|_{L^2(\mathbb{T}^d)} \\
&\lesssim \|\widehat{p}-p\|_{H^{-1}(\mathbb{T}^d)} \|p_\epsilon - p_{\epsilon'}\|_{\mathcal{C}^\beta(\mathbb{T}^d)} \|\widehat{p}-p\|_{H^{-1}(\mathbb{T}^d)} \\
&\lesssim \|\widehat{p}-p\|_{\mathcal{C}^\beta(\mathbb{T}^d)} \|\widehat{p}-p\|_{H^{-1}(\mathbb{T}^d)}^2,
\end{aligned}$$

where the penultimate inequality can be deduced from the stability bound of Proposition 36 and Lemma 59 of Chapter 6. The claim thus follows. □

We are now ready to prove the main result.

7.4.2 Proof of Theorem 30

Fix $\beta \in (0,1)$ so small that $s - \beta/2 > \max\{d/4 - 1, 2\}$. Throughout the proof, we will tacitly make use of the fact that, by Lemma 108, there almost surely exists $N > 0$ such that for all $n \geq N$, there exists $\lambda > 0$ such that

$$\inf_{x \in \mathbb{T}^d} \widehat{p}_n(x) \geq \lambda^{-1}, \quad \text{and} \quad \|\widehat{p}_n\|_{\mathcal{C}^{2+\beta}(\mathbb{T}^d)} \leq \lambda.$$

We will always work over this almost sure event without explicit mention. We also abbreviate $\mathbb{E}_n[\,\cdot\,] := \mathbb{E}[\,\cdot\,|X_1,\dots,X_n]$ and $\mathrm{Var}_n[\,\cdot\,] = \mathrm{Var}[\,\cdot\,|X_1,\dots,X_n]$.

Notice first that

$$\begin{aligned}
&W_2^2(P,Q) - \overline{W}_n \\
&\quad= \int \widehat{\phi}_n d(P - P_n') + \iint \widehat{\Gamma}_n d(P - \widehat{P}_n) d(P - \widehat{P}_n) - \mathcal{U}_n g_n \\
&\quad+ \left[W_2^2(P,Q) - W_2^2(\widehat{P}_n,Q) - \int \widehat{\phi}_n d(P - \widehat{P}_n) - \iint \widehat{\Gamma}_n d(P - \widehat{P}_n) d(P - \widehat{P}_n)\right] \\
&\quad= \int \widehat{\phi}_n d(P - P_n') + \iint \widehat{\Gamma}_n d(P - \widehat{P}_n) d(P - \widehat{P}_n) - \mathcal{U}_n g_n \\
&\quad+ O_p\left(\|\widehat{p}_n - p\|_{\mathcal{C}^\beta(\mathbb{T}^d)} \|\widehat{p}_n - p\|_{H^{-1}(\mathbb{T}^d)}^2\right),
\end{aligned}$$

where we used Theorem 29. By independence of $\mathbf{X}_{(1)}$ and $\mathbf{X}_{(2)}$, it is a simple observation that

$$\sqrt{n}\int \widehat{\phi}_n d(P - P_n') \xrightarrow{d} N(0, \sigma_0^2)$$

as $n \to \infty$, by a similar reasoning as in the proof of Theorem 28. Therefore, to prove the claim, it suffices to show that the remaining terms in the final line of the penultimate display are of order $o_p(n^{-1/2})$. We easily obtain

$$\|\widehat{p}_n - p\|_{\mathcal{C}^\beta(\mathbb{T}^d)} \|\widehat{p}_n - p\|_{H^{-1}(\mathbb{T}^d)}^2 = O_p\left(n^{-\frac{3s+2-\beta}{2s+d}}\right) = o_p(n^{-1/2})$$

under the condition $s - \beta/2 > d/4 - 1$, by Propositions 42–43. It thus remains to show that

$$\mathcal{R}_n := \iint \widehat{\Gamma}_n d(P - \widehat{P}_n) d(P - \widehat{P}_n) - \mathcal{U}_n g_n = o_p(n^{-1/2}). \tag{7.9}$$

To do so, we separately bound the bias and variance of $\mathcal{R}_n$. Regarding the bias, we clearly have

$$\mathbb{E}_n[\mathcal{U}_n g_n] = \iint g_n dP dP = \iint \widetilde{\Gamma}_n d(P - \widehat{P}_n) d(P - \widehat{P}_n),$$

thus, for a large enough constant $C > 0$,

$$\begin{aligned}
\big|\mathbb{E}_n[\mathcal{R}_n]\big| &= \left|\iint \big(\widetilde{\Gamma}_n - \widehat{\Gamma}_n\big) d(P - \widehat{P}_n) d(P - \widehat{P}_n)\right| \\
&= \left|\sum_{j > J_n} \iint \lambda_j^{-1} \eta_j(x) \eta_j(y) d(P - \widehat{P}_n)(x) d(P - \widehat{P}_n)(y)\right| \\
&= \left|\sum_{j > J_n} \lambda_j^{-1} \langle \widehat{p}_n - p, \eta_j \rangle_{L^2(\mathbb{T}^d)}^2\right| \\
&\lesssim \sum_{j > J_n} j^{-2/d} \langle \widehat{p}_n - p, \eta_j \rangle_{L^2(\mathbb{T}^d)}^2.
\end{aligned}$$

Now, reasoning similarly as in the proof of Dunlop et al. (2020, Lemma 17), it holds that

$$\sum_{j > J_n} \lambda_j^s \langle \widehat{p}_n - p, \eta_j \rangle_{L^2(\mathbb{T}^d)}^2 \lesssim \|\widehat{p}_n\|_{H^s(\mathbb{T}^d)}^2 + \|p\|_{H^s(\mathbb{T}^d)}^2 \lesssim 1,$$

thus

$$\big|\mathbb{E}_n[\mathcal{R}_n]\big| \lesssim J_n^{-\frac{2(s+1)}{d}} \sum_{j > J_n} j^{2s/d} \langle \widehat{p}_n - p, \eta_j \rangle_{L^2(\mathbb{T}^d)}^2 \lesssim J_n^{-\frac{2(s+1)}{d}},$$

Let us now turn to the variance. Using an elementary bound on the variance of U-statistics (Robins et al., 2009), we have

$$\mathrm{Var}_n[\mathcal{R}_n] = \mathrm{Var}_n\left[\mathcal{U}_n g_n\right] \lesssim \frac{1}{n}\mathbb{E}_n\left[\left(\int_{\mathbb{T}^d} g_n(X_1', y) dP(y)\right)^2\right] + \frac{1}{n^2}\mathbb{E}_n\left[g_n^2(X_2', X_1')\right].$$

Now, notice that for any $x \in \mathbb{T}^d$,

$$\begin{aligned}\int_{\mathbb{T}^d} g_n(x,y)dP(y) &= \int \widetilde{\Gamma}_n(y,x)d(P-\widehat{P}_n)(y) + \int \left(\int \widetilde{\Gamma}_n(y,x')d(P-\widehat{P}_n)(y)\right) d\widehat{P}_n(x') \\ &\lesssim \|\widehat{p}_n - p\|_{L^\infty(\mathbb{T}^d)},\end{aligned}$$

where we used the fact that $g_n \in L^1(\mathbb{T}^d \times \mathbb{T}^d)$ with uniformly bounded norm. Using this same fact again, we arrive at

$$\mathrm{Var}_n[\mathcal{R}_n] \lesssim \frac{\|\widehat{p}_n - p\|_{L^\infty(\mathbb{T}^d)}}{n} + \frac{1}{n^2}\Big(1 + \mathbb{E}_n\left[\widetilde{\Gamma}_n^2(X_2', X_1')\right]\Big).$$

Regarding the final term, it holds that

$$\begin{aligned}\mathbb{E}\big[\widetilde{\Gamma}_n^2(X_1', X_2')\big] &\lesssim \int\int \left(\sum_{j\leq J_n} \frac{\eta_j(x)\eta_j(y)}{\lambda_j}\right)^2 dP(x)dP(y) \\ &\asymp \sum_{j,k\leq J_n} \lambda_j^{-1}\lambda_k^{-1} \iint \eta_j(x)\eta_j(y)\eta_k(x)\eta_k(y)dxdy \\ &= \sum_{j\leq J_n} \lambda_j^{-2} \asymp J_n^{1-4/d}.\end{aligned}$$

Combining these facts, and recalling that $J_n^{1/d} \asymp n^{\frac{2}{4s+d}}$, we thus arrive at

$$\mathbb{E}[\mathcal{R}_n^2] \lesssim J_n^{-\frac{4(s+1)}{d}} + \frac{\|\widehat{p}_n - p\|_{L^\infty(\mathbb{T}^d)}}{n} + \frac{J_n^{1-4/d}}{n^2} \lesssim n^{-\frac{8(s+1)}{4s+d}} + o(n^{-1}).$$

Under the condition $s > d/4 - 1$, the above display is of order $o(n^{-1})$. The claim follows from here. □

7.A Proofs of Central Limit Theorems for Plugin Estimators

The aim of this appendix is to prove Theorem 26. We also state and prove Proposition 40, regarding the question of variance estimation. We begin by deriving limit laws for the functional $\int \phi_0(\widehat{p}_n - p)$, which form an important component of our central limit theorems. Here, $\widehat{p}_n$ is one of the estimators $\widehat{p}_n^{(\mathrm{bc})}$, and $\widehat{p}_n^{(\mathrm{ker})}$, which respectively arise from the boundary-corrected and kernel density estimators $\widetilde{p}_n^{(\mathrm{bc})}$, and $\widetilde{p}_n^{(\mathrm{ker})}$. We also write

$$p_{J_n}^{(\mathrm{bc})} = \mathbb{E}[\widetilde{p}_n^{(\mathrm{bc})}], \quad p_{h_n}^{(\mathrm{ker})} = \mathbb{E}[\widetilde{p}_n^{(\mathrm{ker})}].$$

We have the following.

Lemma 89. *Let $\epsilon, s > 0$, and let $h_n^{-1} \asymp 2^{J_n} \uparrow \infty$.*

(i) *(Unit Hypercube) Let* $p \in \mathcal{C}^\epsilon([0,1]^d)$ *be positive over* $[0,1]^d$. *Assume that* $\phi_0 \in \mathcal{C}^s([0,1]^d)$ *satisfies* $\mathrm{Var}_P[\phi_0(X)] > 0$. *Then, as* $n \to \infty$,

$$\sqrt{n}\int \phi_0(\widehat{p}_n^{(\mathrm{bc})} - p_{J_n}^{(\mathrm{bc})}) \rightsquigarrow N(0, \mathrm{Var}_P[\phi_0(X)]).$$

(ii) *(Flat Torus) Let* $p \in \mathcal{C}^\epsilon(\mathbb{T}^d)$ *be positive over* $\mathbb{T}^d$. *Assume that* $\phi_0 \in \mathcal{C}^s(\mathbb{T}^d)$ *satisfies* $\mathrm{Var}_P[\phi_0(X)] > 0$. *Then, as* $n \to \infty$,

$$\sqrt{n}\int \phi_0(\widehat{p}_n^{(\mathrm{ker})} - p_{h_n}^{(\mathrm{ker})}) \rightsquigarrow N(0, \mathrm{Var}_P[\phi_0(X)]).$$

7.A.1 Proof of Lemma 89

The proof is standard, thus we only prove claim (i). The remaining claims can be proven similarly. For simplicity, we write $\Psi_{j_0-1}^{\mathrm{bc}} = \Phi^{\mathrm{bc}}$ throughout the proof. Reasoning as in the proof of Lemma 36, and in particular using Lemma 105, it holds that

$$\begin{aligned}\sqrt{n}\int \phi_0(\widehat{p}_n^{(\mathrm{bc})} - p_{J_n}^{(\mathrm{bc})}) &= \sqrt{n}\int \phi_0(\widetilde{p}_n^{(\mathrm{bc})} - p_{J_n}^{(\mathrm{bc})}) + \sqrt{n}\int \phi_0(\widehat{p}_n^{(\mathrm{bc})} - \widetilde{p}_n^{(\mathrm{bc})}) \\ &= \sqrt{n}\int \phi_0(\widetilde{p}_n^{(\mathrm{bc})} - p_{J_n}^{(\mathrm{bc})}) + o_p(1) \\ &= \frac{1}{\sqrt{n}}\sum_{i=1}^n (Z_{n,i} - \mathbb{E}[Z_{n,i}]),\end{aligned}$$

where we write

$$Z_{n,i} = \sum_{j=j_0-1}^{J_n}\sum_{\xi\in\Psi_j^{\mathrm{bc}}} \xi(X_i)\gamma_\xi, \quad i = 1,\dots,n,$$

and where $\gamma_\xi = \int \phi_0\xi$ for all $\xi \in \Psi^{\mathrm{bc}}$. By Lyapunov's central limit theorem (Billingsley (1968), Theorem 7.3), it holds that

$$\frac{1}{\sqrt{\sum_{i=1}^n \mathrm{Var}[Z_{n,i}]}}\sum_{i=1}^n (Z_{n,i} - \mathbb{E}[Z_{n,i}]) \rightsquigarrow N(0,1), \tag{7.10}$$

provided that for some $p > 2$,

$$\frac{\sum_{i=1}^n \mathbb{E}\left[|Z_{n,i} - \mathbb{E}Z_{n,i}|^p\right]}{\left(\sum_{i=1}^n \mathrm{Var}[Z_{n,i}]\right)^{p/2}} \to 0. \tag{7.11}$$

Now, using Lemma 101, it holds that

$$\sup_{n\geq 1}\sup_{1\leq i\leq n}|Z_{n,i}| \leq \sup_{n\geq 1}\left\|\sum_{j=j_0-1}^{J_n}\sum_{\xi\in\Psi_j^{\mathrm{bc}}}\xi\gamma_\xi\right\|_\infty$$

$$\begin{aligned}
&\leq \sum_{j=j_0-1}^{\infty} \|(\gamma_\xi)_{\xi\in\Psi_j^{\mathrm{bc}}}\|_{\ell_\infty} \left(\sup_{\xi\in\Psi_j^{\mathrm{bc}}} \|\xi\|_\infty\right) \left\| \sum_{\xi\in\Psi_j^{\mathrm{bc}}} I(|\xi|>0) \right\|_\infty \\
&\lesssim \sum_{j=j_0-1}^{\infty} 2^{-j\left(\frac{d}{2}+s\right)} 2^{\frac{dj}{2}} \lesssim \sum_{j=j_0-1}^{\infty} 2^{-js} < \infty. \qquad (7.12)
\end{aligned}$$

On the other hand, under the stated conditions, it follows from Lemma 36 that

$$\sum_{i=1}^{n} \mathrm{Var}[Z_{n,i}] = n(\mathrm{Var}_P[\phi_0(X)] + o(1)). \qquad (7.13)$$

Since $\mathrm{Var}_P[\phi_0(X)] > 0$, the denominator in equation (7.11) is of the order $n^{p/2}$, while the numerator is of order n by equation (7.12). It follows that Lyapunov's condition (7.11) holds for all $p > 2$. The claim thus follows from equations (7.10) and (7.13). □

7.A.2 Proof of Theorem 26

Assume first that $\sigma_0, \sigma_1 > 0$. We begin with part (i). Under the stated conditions on the densities, it follows from Theorem 4 that φ_0 satisfies condition **A1(λ)** for some $\lambda > 0$. Apply the stability bound of Theorem 18 to obtain,

$$0 \leq W_2^2(\widehat{P}_n^{(\mathrm{ker})}, Q) - W_2^2(P,Q) - \int \phi_0 d(\widehat{P}_n^{(\mathrm{ker})} - P) \leq W_2^2(\widehat{P}_n^{(\mathrm{ker})}, P).$$

Using the convergence rate of $\widehat{P}_n^{(\mathrm{ker})}$ under W_2^2 in Proposition 46, and Lemma 47 regarding the bias of $\int \phi_0 d\widehat{P}_n^{(\mathrm{ker})}$, we obtain

$$W_2^2(\widehat{P}_n^{(\mathrm{ker})}, Q) - W_2^2(P,Q) = \int \phi_0(\widehat{p}_n^{(\mathrm{ker})} - p_{h_n}^{(\mathrm{ker})}) + O_p\left(n^{-\frac{2\alpha}{2(\alpha-1)+d}} \vee \frac{(\log n)^2}{n}\right).$$

Using the assumption $2(\alpha+1) > d$, deduce that

$$\sqrt{n}\left(W_2^2(\widehat{P}_n^{(\mathrm{ker})}, Q) - W_2^2(P,Q)\right) = \sqrt{n}\int \phi_0(\widehat{p}_n^{(\mathrm{ker})} - p_{h_n}^{(\mathrm{ker})}) + o_p(1).$$

Apply Lemma 89 to deduce that

$$\sqrt{n}\left(W_2^2(\widehat{P}_n^{(\mathrm{ker})}, Q) - W_2^2(P,Q)\right) \rightsquigarrow N(0,\sigma_0^2), \quad \text{as } n \to \infty.$$

By the same reasoning, but now using the two-sample stability bound of Proposition 23, we also have

$$\begin{aligned}
&\sqrt{\frac{nm}{n+m}}\left(W_2^2(\widehat{P}_n^{(\mathrm{ker})}, Q_m^{(\mathrm{ker})}) - W_2^2(P,Q)\right) \\
&\qquad = \sqrt{(1-\rho)n}\int \phi_0(\widehat{p}_n^{(\mathrm{ker})} - p_{h_n}^{(\mathrm{ker})}) + \sqrt{\rho m}\int \psi_0(\widehat{q}_m^{(\mathrm{ker})} - q_{h_m}^{(\mathrm{ker})}) + o_p(1),
\end{aligned}$$

as $n, m \to \infty$ such that $n/(n+m) \to \rho \in [0,1]$. By Lemma 89 and the independence of $X_1, \dots, X_n, Y_1, \dots, Y_m$, we deduce that

$$\sqrt{\frac{nm}{n+m}}\Big(W_2^2(\widehat{P}_n^{(\mathrm{ker})}, Q_m^{(\mathrm{ker})}) - W_2^2(P,Q)\Big) \rightsquigarrow N(0, \sigma_\rho^2),$$

as $n, m \to \infty$ such that $n/(n+m) \to \rho$. This proves claim (i) for kernel estimators. Claim (ii) regarding the boundary-corrected wavelet estimators $(\widehat{P}_n, \widehat{Q}_m)$ now follows analogously by using Lemma 36 to bound the bias of $\int \phi_0 d\widehat{P}_n$, Lemma 106 to bound the convergence rate of $\widehat{P}_n$ in Wasserstein distance, Proposition 28, to bound the bias of the plugin estimator of the Wasserstein distance, and Lemma 89 to obtain the limiting distribution of $\int \phi_0(\widetilde{p}_n - \mathbb{E}[\widetilde{p}_n])$.

Finally, to prove part (v), apply Corollary 19 and the result of Divol (2021) to deduce that, for $\Omega \in \{[0,1]^d, \mathbb{T}^d\}$, since the densities p and q are bounded and bounded away from zero, we have

$$\sqrt{n}W_2^2(P_n, P) = o_p(1), \quad \sqrt{m}W_2^2(Q_m, Q) = o_p(1),$$

as $n, m \to \infty$, whenever $d \le 3$. Therefore, using Theorem 18, Proposition 23, and Proposition 27, we obtain

$$\begin{aligned}\sqrt{n}\Big(W_2^2(P_n, Q) - W_2^2(P,Q)\Big) &= \sqrt{n}\int \phi_0 d(P_n - P) + o_p(1),\\ \sqrt{\frac{nm}{n+m}}\Big(W_2^2(P_n, Q_m) - W_2^2(P,Q)\Big) &= \sqrt{(1-\rho)n}\int \phi_0 d(P_n - P)\\ &\quad + \sqrt{\rho m}\int \psi_0 d(Q_m - Q) + o_p(1).\end{aligned}$$

Claim (v) then follows by the classical central limit theorem.

It thus remains to consider the situation where $\sigma_1 = 0$ or $\sigma_0 = 0$. Notice that the Kantorovich potentials ϕ_0 and ψ_0 are almost everywhere constant if and only if $P = Q$. As a result, the statements "$\sigma_0 = 0$", "$\sigma_1 = 0$", and "$P = Q$" are equivalent, thus it remains to prove the claim when $P = Q$. In this case, it suffices to show that $\sqrt{n}W_2^2(\widehat{P}_n, P) = o_p(1)$ and $\sqrt{\frac{nm}{n+m}}W_2^2(\widehat{P}_n, \widehat{Q}_m) = o_p(1)$ for the various estimators $\widehat{P}_n$ and $\widehat{Q}_m$ under consideration. But these assertions are a direct consequence of the aforementioned convergence rates of these estimators in Wasserstein distance, under the assumptions of each of parts (i)–(v). The claim thus follows. □

7.A.3 Variance Estimation

We now state a simple result regarding the estimation of variances appearing in Theorem 26. In what follows, let Ω be equal to $\mathbb{T}^d$, or to a compact and connected subset of $\mathbb{R}^d$, and let $P, Q \in \mathcal{P}_{\mathrm{ac}}(\Omega)$. Let (ϕ_0, ψ_0) be a pair of Kantorovich potentials in the optimal transport problem from P to Q. Furthermore, let

$$X_1, \dots, X_n \sim P, \quad Y_1, \dots, Y_m \sim Q$$

be i.i.d. samples which are independent of each other, and let P_n and Q_m denote their respective empirical measures.

Proposition 40. Let the distributions $P, Q \in \mathcal{P}_{\mathrm{ac}}(\Omega)$ have positive densities over Ω. Let $\widehat{P}_n$ and $\widehat{Q}_m$ be estimators such that

$$W_2(\widehat{P}_n, P) = o_p(1), \quad \text{and} \quad W_2(\widehat{Q}_m, Q) = o_p(1),$$

as $n, m \to \infty$. Let $(\widehat{\phi}_{nm}, \widehat{\psi}_{nm})$ be a uniformly bounded pair of Kantorovich potentials in the optimal transport problem from $\widehat{P}_n$ to $\widehat{Q}_m$. Then, as $n, m \to \infty$,

$$\begin{aligned}
\widehat{\sigma}^2_{0,nm} &:= \mathrm{Var}_{U\sim P_n}[\widehat{\phi}_{nm}(U)] \xrightarrow{p} \mathrm{Var}[\phi_0(X)], \text{ and,} \\
\widehat{\sigma}^2_{1,nm} &:= \mathrm{Var}_{V\sim Q_m}[\widehat{\psi}_{nm}(V)] \xrightarrow{p} \mathrm{Var}[\psi_0(Y)].
\end{aligned}$$

Note that the assumption of uniform boundedness of the fitted potentials can always be satisfied, due to the compactness of Ω (Villani, 2003, Remark 1.13). In particular, the conditions of Proposition 40 are met for any of the estimators $(\widehat{P}_n, \widehat{Q}_m)$ appearing in the statement of Theorem 26.

Proof of Proposition 40. To prove the claim, we shall make use of the following stability result for Kantorovich potentials over compact metric spaces, due to Santambrogio (2015, Theorem 1.52), which we only state in the generality required for our proofs.

Lemma 90 (Santambrogio (2015)). *Let $P, Q \in \mathcal{P}_{\mathrm{ac}}(\Omega)$, and assume that at least one of P and Q has support equal to Ω. Let $(\overline{P}_k)_{k\geq 1}, (\overline{Q}_k)_{k\geq 1} \subseteq \mathcal{P}(\Omega)$ be sequences which respectively converge to P, Q weakly. Let (ϕ_k, ψ_k) denote a pair of Kantorovich potentials in the optimal transport problem from $\overline{P}_k$ to $\overline{Q}_k$, for all $k \geq 1$. Then, it holds that $\phi_k \to \phi_0$ and $\psi_k \to \psi_0$ as $k \to \infty$, the convergence being uniform over Ω, for some pair of Kantorovich potentials (ϕ_0, ψ_0) in the optimal transport problem from P to Q, which is uniquely defined up to translation by constants.*

We will prove the claim for $\widehat{\sigma}^2_{0,nm}$, and a symmetric argument may be used for $\widehat{\sigma}^2_{1,nm}$. By Lemma 90, there exists a random variable a_{nm} such that $\widehat{\phi}^o_{nm} := \widehat{\phi}_{nm} + a_{nm}$ and $\widehat{\psi}^o_{nm} := \widehat{\psi}_{nm} - a_{nm}$ converge uniformly to ϕ_0 and ψ_0 respectively. We have,

$$\begin{aligned}
|\widehat{\sigma}^2_{0,nm} - \sigma^2_0| &= \Big|\mathrm{Var}_{P_n}[\widehat{\phi}_{nm}(U)] - \mathrm{Var}_P[\phi_0(X)]\Big| \\
&= \Big|\mathrm{Var}_{P_n}[\widehat{\phi}^o_{nm}(U)] - \mathrm{Var}_P[\phi_0(X)]\Big| \\
&\lesssim \left|\int (\widehat{\phi}^o_{nm})^2 dP_n - \int_{\mathbb{T}^d} \phi_0^2 dP\right| + \left|\int \widehat{\phi}^o_{nm} dP_n - \int_{\mathbb{T}^d} \phi_0 dP\right| \\
&\lesssim \left|\int (\widehat{\phi}^o_{nm})^2 d(P_n - P)\right| + \left|\int \widehat{\phi}^o_{nm} d(P_n - P)\right| + \big\|\widehat{\phi}^o_{nm} - \phi_0\big\|_{L^2(P)}.
\end{aligned}$$

Since $\widehat{\phi}^o_{nm}$ is convex up to translation by a quadratic function, and uniformly bounded, it must be Lipschitz with respect to $\|\cdot\|$ over the compact set Ω, with a uniform constant depending only on the diameter of this set (Hiriart-Urruty and Lemaréchal (2004), Lemma 3.1.1, p. 102). Thus, $(\widehat{\phi}^o_{nm})^2$ is also Lipschitz over Ω with uniform constant. The set of Lipschitz functions with a uniformly bounded Lipschitz constant, over any given compact domain, forms a Glivenko-Cantelli class (van der Vaart and Wellner (1996), Theorem 2.7.1), thus the first two terms on the right-hand side of the above display vanish in probability. The final term vanishes due to the uniform convergence of $\widehat{\phi}^o_{nm}$ to ϕ_0. □

7.B Proofs of Efficiency Lower Bounds

Throughout this appendix, given $Q \in \mathcal{P}_{\mathrm{ac}}(\Omega)$, we abbreviate the functional Φ_Q by Φ, and the influence functions $\widetilde{\Phi}_{(P,Q)}$ and $\widetilde{\Psi}_{(P,Q)}$ by $\widetilde{\Phi}$ and $\widetilde{\Psi}$, respectively.

We begin by defining the differentiable paths $(P_{t,h_1})_{t\geq 0}$ and $(Q_{t,h_2})_{t\geq 0}$, for all $(h_1, h_2) \in \mathbb{R}^2$, as announced in Section 7.2. We follow a construction from Example 1.12 of van der Vaart (2002). Recall that $P, Q \in \mathcal{P}_{\mathrm{ac}}(\Omega)$ admit respective densities p, q. Let $\zeta \in \mathcal{C}^\infty(\mathbb{R}) \cap \mathcal{C}^2(\mathbb{R})$ be a bounded nonnegative map, which is bounded away from zero over $\mathbb{R}$ by a positive constant, and which satisfies $\zeta(0) = \zeta'(0) = \zeta''(0) = 1$. For any functions $f \in L^2_0(P)$ and $g \in L^2_0(Q)$, define $P^f_t, Q^g_t \in \mathcal{P}_{\mathrm{ac}}(\Omega)$ to be the distributions with densities

$$p^f_t(x) \propto \zeta(tf(x))p(x), \quad q^g_t(y) \propto \zeta(tg(y))q(y), \tag{7.14}$$

for all $x, y \in \Omega$ and $t \geq 0$. Since ζ is bounded away from zero over $\mathbb{R}$, notice that the implicit normalizing constants in the above display are bounded from above by a constant which does not depend on t, f, g, p, q. We now turn to the proofs of Lemma 86 and Theorem 27.

7.B.1 Proof of Lemma 86

Let $f \in \dot{\mathcal{P}}_P$ be an arbitrary score function, and abbreviate the differentiable path $P_t := P^f_t$, and its density $p_t := p^f_t$, for all $t \geq 0$. Here, we use the definition in equation (7.14). Let (ϕ_t, ψ_t) denote a pair of Kantorovich potentials in the optimal transport problem from P_t to Q, which we may and do choose to be uniformly bounded by $\mathrm{diam}(\Omega)^2$, and hence uniformly bounded in t. By the Kantorovich duality, one has

$$\begin{aligned}\Phi(P_t) - \Phi(P) &= \sup_{(\phi,\psi)\in\mathcal{K}} \left[\int \phi dP_t + \int \psi dQ\right] - \int \phi_0 dP - \int \psi_0 dQ \\ &\geq \int \phi_0 dP_t + \int \psi_0 dQ - \int \phi_0 dP - \int \psi_0 dQ = \int \phi_0 d(P_t - P), \\ \Phi(P_t) - \Phi(P) &= \int \phi_t dP_t + \int \psi_t dQ - \sup_{(\phi,\psi)\in\mathcal{K}} \left[\int \phi dP - \int \psi dQ\right] \leq \int \phi_t d(P_t - P).\end{aligned} \tag{7.15}$$

By construction, the map $t \in [0,\infty) \mapsto p_t(x)$ is differentiable for every $x \in \Omega$, and letting $\Delta_t(x) = (p_t(x) - p(x))/t$, we have

$$\lim_{t\to 0} \Delta_t(x) = \frac{\partial}{\partial t} p_t(x)\bigg|_{t=0} = f(x)p(x). \tag{7.16}$$

Now, notice that for all $t \geq 0$,

$$|\Delta_t(x)| \lesssim \left| \frac{\zeta(tf(x)) - 1}{t} \right| p(x) = \left| \frac{\zeta(tf(x)) - \zeta(0)}{t} \right| p(x) \leq \|\zeta\|_{\mathcal{C}^1(\mathbb{R}^d)} f(x) p(x).$$

Since $f \in L_0^2(P) \subseteq L_0^1(P)$, we deduce that $\Delta_t(x)$ is dominated by an integrable function, uniformly in t. Since ϕ_0 is uniformly bounded, we also deduce that the map $|\phi_0||\Delta_t - fp|$ is dominated by an integrable function. We then have, by equation (7.16) and the Dominated Convergence Theorem,

$$\begin{aligned}
\liminf_{t \to 0} \frac{\Phi(P_t) - \Phi(P)}{t} &\geq \liminf_{t \to 0} \int_\Omega \phi_0 \Delta_t d\mathcal{L} \\
&= \int_\Omega \phi_0 f dP + \liminf_{t \to 0} \int_\Omega \phi_0 \left[\Delta_t - fp\right] d\mathcal{L} \\
&\geq \int_\Omega \phi_0 f dP - \limsup_{t \to 0} \int_\Omega |\phi_0| \left|\Delta_t - fp\right| d\mathcal{L} \\
&\geq \int_\Omega \phi_0 f dP - \int_\Omega |\phi_0| \limsup_{t \to 0} \left|\Delta_t - fp\right| d\mathcal{L} = \int \phi_0 f dP, \qquad (7.17)
\end{aligned}$$

and similarly,

$$\begin{aligned}
\limsup_{t \to 0} \frac{\Phi(P_t) - \Phi(P)}{t} &\leq \limsup_{t \to 0} \int \phi_t \Delta_t d\mathcal{L} \\
&\leq \limsup_{t \to 0} \int \phi_t f dP + \left(\sup_{t \geq 0} \|\phi_t\|_{L^\infty(\Omega)} \right) \limsup_{t \to 0} \int \left|\Delta_t - fp\right| d\mathcal{L} \\
&= \limsup_{t \to 0} \int \phi_t f dP.
\end{aligned}$$

Let $t_k \downarrow 0$ be a sequence achieving the limit superior, in the sense that $\lim_{k \to \infty} \int \phi_{t_k} f dP = \limsup_{t \to 0} \int \phi_t dP$. Up to taking a subsequence of (t_k), Lemma 90 implies that ϕ_{t_k} converges uniformly to a Kantorovich potential f_0 from P to Q, which is unique up to translation by a constant, and which therefore takes the form $f_0 = \phi_0 + a$ for some $a \in \mathbb{R}$. The limit superior clearly continues to be achieved along this subsequence, thus we replace it by (ϕ_{t_k}) without loss of generality. We thus have

$$\limsup_{t \to 0} \int \phi_t f dP = \lim_{k \to \infty} \int \phi_{t_k} f dP = \int \left(\lim_{k \to \infty} \phi_{t_k} \right) f dP = \int (\phi_0 + a) f dP = \int \phi_0 f dP,$$

where the interchange of limit and integration holds again by the Dominated Convergence Theorem, since ϕ_t are uniformly bounded, and $f \in L_0^2(P)$. Combine this fact with equation (7.17) to deduce that

$$\lim_{t \to 0} \frac{\Phi(P_t) - \Phi(P)}{t} = \int \phi_0 f dP.$$

It follows that $\widetilde{\Phi} = \phi_0 - \int \phi_0 dP$ is an influence function of Φ with respect to $\dot{\mathcal{P}}_P$. Since we assumed that $\widetilde{\Phi} \in \dot{\mathcal{P}}_P$, it must in fact be the case that $\widetilde{\Phi}$ is the unique efficient influence function of Φ with respect to $\dot{\mathcal{P}}_P$ (van der Vaart, 2002), and the claim follows. □

7.B.2 Proof of Theorem 27

We shall use the following abbrevations of the differentiable paths defined in equation (7.14). For any $h \in \mathbb{R}$ and $t \geq 0$, if $f = h\widetilde{\Phi}$ and $g = h\widetilde{\Psi}$, we write

$$P_{t,h} := P_t^f, \quad p_{t,h}(x) := p_t^f(x) = c_h(t)\zeta(th\widetilde{\Phi}(x))p(x), \tag{7.18}$$

$$Q_{t,h} := Q_t^g, \quad q_{t,h}(y) := q_t^g(y) = k_h(t)\zeta(th\widetilde{\Psi}(y))q(y), \tag{7.19}$$

for all $x, y \in \mathbb{T}^d$, where the normalizing constants are explicitly denoted

$$c_h(t) = \left(\int_{\mathbb{T}^d} \zeta(th\widetilde{\Phi}(x))dP(x)\right)^{-1}, \quad k_h(t) = \left(\int_{\mathbb{T}^d} \zeta(th\widetilde{\Psi}(y))dQ(y)\right)^{-1}.$$

In this case, the collections $\{(P_{t,h})_{t\geq 0} : h \in \mathbb{R}\}$ and $\{(Q_{t,h})_{t\geq 0} : h \in \mathbb{R}\}$ respectively have score functions given by the tangent spaces

$$\dot{\mathcal{P}}_P = \{h\widetilde{\Phi} : h \in \mathbb{R}\}, \quad \dot{\mathcal{P}}_Q = \{h\widetilde{\Psi} : h \in \mathbb{R}\}.$$

We begin by showing that there exist $\overline{M}, \bar{\gamma}, \bar{u} > 0$ such that $p_{t,h} \in \mathcal{C}^{\alpha-1}(\mathbb{T}^d; \overline{M}, \bar{\gamma})$ uniformly in $t|h| \leq \bar{u}$. An identical argument may then be used to show that $q_{t,h} \in \mathcal{C}^{\alpha-1}(\mathbb{T}^d; \overline{M}, \bar{\gamma})$ for all appropriate t, h. Our proof then proceeds by proving parts (i) and (ii).

Since $p \geq \gamma^{-1}$, and since ζ is bounded from below by a positive constant, it is clear that there must exist $\bar{\gamma} > 0$ depending on γ and ζ such that

$$\bar{\gamma}^{-1} \leq p_{t,h} \quad \text{over } \mathbb{T}^d, \text{ for all } t \geq 0, h \in \mathbb{R}. \tag{7.20}$$

We next prove the uniform Hölder continuity of $p_{t,h}$. We begin by studying the Hölder continuity of the map $\zeta(th\widetilde{\Phi}(\cdot))$. Since $p, q \in \mathcal{C}^{\alpha-1}(\mathbb{T}^d; M, \gamma)$, and since we assumed $\alpha \notin \mathbb{N}$, we have by Theorem 4 that, for some constant $\lambda > 0$ depending only on M, γ, α,

$$\|\widetilde{\Phi}\|_{\mathcal{C}^{\alpha+1}(\mathbb{T}^d)} \leq \lambda. \tag{7.21}$$

Furthermore, recall that $\zeta \in \mathcal{C}^\infty(\mathbb{R})$. Thus, by the multivariate Faà di Bruno formula (see, for instance, Encinas and Masque (2003); Constantine and Savits (1996)), it holds that for all multi-indices $1 \leq |\beta| \leq \lfloor \alpha + 1 \rfloor$,

$$D^\beta \zeta(th\widetilde{\Phi}(\cdot)) = \beta! \sum_{\ell=0}^{|\beta|} (th)^\ell \zeta^{(\ell)}(th\widetilde{\Phi}(\cdot)) \sum_{(e_j),(\tau_j)} \prod_{j=1}^d \frac{1}{e_j!} \left(\frac{1}{\tau_j!} D^{\tau_j}\widetilde{\Phi}(\cdot)\right)^{e_j},$$

where the second summation is taken over all indices $(e_j)_{1\leq j\leq d} \subseteq \mathbb{N}$ and multi-indices $(\tau_j)_{1\leq j\leq d} \subseteq \mathbb{N}^d$ such that for some $1 \leq s \leq d-1$, $\tau_j = 0$ and $e_j = 0$ for all $1 \leq j \leq s$, $e_j \neq 0$ and $\tau_j \neq 0$ for $s+1 \leq j \leq d$, and for which it holds that $\sum_{j=1}^d e_\tau = \ell$ and $\sum_{j=1}^d e_j\tau_j = \beta$. Furthermore, $\beta! = \beta_1! \dots \beta_d!$, and $\tau_j!$ is defined similarly for all j. Since $\zeta \in \mathcal{C}^\infty(\mathbb{R})$, its derivatives of all orders less than $\alpha + 1$ are uniformly bounded over any fixed compact set. Since $\widetilde{\Phi}$ is bounded, we deduce that for any $\overline{u} > 0$,

$$\sup_{0 \leq \ell \leq \lfloor \alpha+1 \rfloor} \sup_{\substack{t \geq 0, h \in \mathbb{R} \\ t|h| \leq \overline{u}}} \|(th)^{\ell} \zeta^{(\ell)}(th\widetilde{\Phi}(\cdot))\|_{\infty} \lesssim_{\overline{u},\alpha} 1.$$

Furthermore, we have $\|D^{\tau}\widetilde{\Phi}\|_{\infty} \leq \lambda$ for any $0 \leq |\tau| \leq \lfloor \alpha + 1 \rfloor$. This fact together with the preceding two displays implies

$$\sup_{0 \leq |\beta| \leq \lfloor \alpha+1 \rfloor} \sup_{\substack{t \geq 0, h \in \mathbb{R} \\ t|h| \leq \overline{u}}} \|D^{\beta} \zeta(th\widetilde{\Phi}(\cdot))\|_{\infty} \lesssim_{\lambda,\overline{u},\alpha} 1. \tag{7.22}$$

Now, recall that $p_{t,h}(\cdot) = c_h(t)\zeta(th\widetilde{\Phi}(\cdot))p(\cdot)$, and that $c_h(t)$ is uniformly bounded in h and t because ζ is bounded away from zero by a positive constant. Thus, using the above display, the fact that $p \in \mathcal{C}^{\alpha-1}(\mathbb{T}^d; M)$, and Lemma 95, we deduce there exists a constant $\overline{M} > 0$, depending only on $M, \overline{u}, \alpha$ and the choice of ζ, such that

$$\sup_{\substack{t \geq 0, h \in \mathbb{R} \\ t|h| \leq \overline{u}}} \|p_{t,h}\|_{\mathcal{C}^{\alpha-1}(\mathbb{T}^d)} \leq \overline{M}.$$

Combine this fact with equation (7.20) to deduce that

$$p_{t,h} \in \mathcal{C}^{\alpha-1}(\mathbb{T}^d; \overline{M}, \overline{\gamma}), \quad \text{for all } t \geq 0, h \in \mathbb{R}, t|h| \leq \overline{u}.$$

We now prove part (i). Since $\widetilde{\Phi} \in \dot{\mathcal{P}}_P$, it follows from Lemma 86 that $\widetilde{\Phi}$ is the efficient influence function of Φ relative to $\dot{\mathcal{P}}_P$. Since $\dot{\mathcal{P}}_P$ is a vector space, it follows from Theorem 25.21 of van der Vaart (1998) that for any estimator sequence U_n,

$$\sup_{\substack{\mathcal{I} \subseteq \mathbb{R} \\ |\mathcal{I}| < \infty}} \liminf_{n \to \infty} \sup_{h \in \mathcal{I}} n\mathbb{E}_{n,h}\big|U_n - \Phi_Q(P_{n^{-1/2},h})\big|^2 \geq \mathrm{Var}_P[\phi_0(X)],$$

where the infimum is over all estimator sequences.

We next prove part (ii). Inspired by the proof of Theorem 11 of Berrett and Samworth (2023), our goal will be to invoke a more general version of Theorem 25.21 of van der Vaart (1998), given in Theorem 3.11.5 of van der Vaart and Wellner (1996), whose statement we briefly summarize here. Let H be a Hilbert space with inner product $\langle \cdot, \cdot \rangle_H$, and norm $\|\cdot\|_H$. Let $(\mathcal{X}_n, \mathcal{A}_n, \mu_{n,h} : h \in H)$ be a sequence of asymptotically normal experiments (as defined in Section 3.11 of van der Vaart and Wellner (1996)). A parameter sequence $(\kappa_n(h) : h \in H) \subseteq \mathbb{R}$ is said to be regular if there exists a nonnegative sequence (r_n) such that

$$r_n\big(\kappa_n(h) - \kappa_n(0)\big) \to \dot{\kappa}(h), \quad h \in H,$$

for a continuous linear map $\dot{\kappa} : H \to \mathbb{R}$. Denote by $\dot{\kappa}^* : \mathbb{R} \to H$ the adjoint of $\dot{\kappa}$, namely the map satisfying $\langle \dot{\kappa}^*(b^*), h \rangle_H = b^* \dot{\kappa}(h)$ for all $h \in H$.

Lemma 91 (van der Vaart and Wellner (1996), Theorem 3.11.5)**.** *Let the sequence of experiments* $(\mathcal{X}_n, \mathcal{A}_n, \mu_{n,h} : h \in H)$ *be asymptotically normal, and let the parameter sequence* $(\kappa_n(h) : h \in H)$ *be regular. Suppose there exists a Gaussian random variable* G *such that for all* $b^* \in \mathbb{R}$, $b^* G \sim N(0, \|\dot{\kappa}^*(b^*)\|_H^2)$*. Then, for any estimator sequence* $(U_n)_{n\geq 1}$*,*

$$\sup_{\substack{\mathcal{I}\subseteq H\\ |\mathcal{I}|<\infty}} \liminf_{n\to\infty} \sup_{h\in\mathcal{I}} r_n^2 \mathbb{E}_{\mu_{n,h}}(U_n - \kappa_n(h))^2 \geq \mathrm{Var}[G].$$

Returning to the proof, define the Hilbert space $H = \mathbb{R}^2$ with inner product

$$\langle (h_1,h_2),(h_1',h_2')\rangle_H = h_1 h_1' \operatorname{Var}_P[\phi_0(X)] + h_2 h_2' \operatorname{Var}_Q[\psi_0(Y)], \quad (h_1,h_2),(h_1',h_2') \in H,$$

and the sequence of experiments

$$\mu_{n,h} = P_{n^{-1/2},h_1}^{\otimes n} \otimes Q_{m^{-1/2},h_2}^{\otimes m}, \quad h = (h_1,h_2) \in \mathbb{R}^2,$$

endowed with the standard Borel σ-algebra. Here, m is viewed as a function of n which satisfies $n/(n+m) \to \rho \in [0,1]$ as $n \to \infty$. The following result can be deduced from Section 7.5 of Berrett and Samworth (2023) with minor modifications, using our assumptions placed on ζ.

Lemma 92 (Berrett and Samworth (2023))**.** *The sequence of experiments* $(\mu_{n,h} : h \in H)$ *is asymptotically normal.*

For all $h = (h_1,h_2) \in H$, let $\kappa_n(h) = \Psi(P_{n^{-1/2},h_1}, Q_{m^{-1/2},h_2})$, where again m is treated as a function of n. Notice that $\kappa_n(0) = \Psi(P,Q)$ for any $n \geq 1$. By following the same argument as in the proof of Lemma 86, using the Kantorovich duality and the stability result for Kantorovich potentials in Lemma 90, one has

$$\kappa_n(h) - \kappa_n(0) = \int \phi_0 d(P_{n^{-1/2},h_1} - P) + \int \psi_0 d(Q_{m^{-1/2},h_2} - Q) + o(1).$$

Now, since $\zeta(0) = \zeta'(0) = 1$ and $\int \widetilde{\Phi} dP = 0$, we have for all $t \geq 0$,

$$\left|\frac{1}{c_{h_1}(t)} - 1\right| = \left|\int \left[\zeta(th_1\widetilde{\Phi}(x)) - 1 - th_1\widetilde{\Phi}(x)\right] dP(x)\right| \lesssim \|\zeta\|_{\mathcal{C}^2(\mathbb{R})} t^2 h_1^2 \|\widetilde{\Phi}\|_{L^2(P)}.$$

Recall that p and ζ are bounded, and that $c_{h_1}(n^{-1/2})$ is uniformly bounded in h_1 and n, thus for all $x \in \mathbb{T}^d$,

$$\begin{aligned} p_{n^{-1/2},h_1}(x) - p(x) &= p(x)\left[c_{h_1}(n^{-1/2})\zeta\big(h_1 n^{-1/2}\widetilde{\Phi}(x)\big) - 1\right] \\ &= p(x)\left[\zeta\big(h_1 n^{-1/2}\widetilde{\Phi}(x)\big) - 1\right] + O\left(\frac{\|p\|_\infty \|\zeta\|_\infty h_1^2}{n}\right) \\ &= p(x) h_1 n^{-1/2}\widetilde{\Phi}(x) + O\left(h_1^2/n\right), \end{aligned}$$

where we again used the fact that $\zeta(0) = \zeta'(0) = 1$. Similarly, for all $y \in \mathbb{T}^d$,

$$q_{m^{-1/2},h_2}(y) - q(y) = q(y) h_2 m^{-1/2}\widetilde{\Psi}(y) + O\left(h_2^2/m\right),$$

implying that,

$$\begin{aligned}\kappa_n(h) - \kappa_n(0) &= h_1 n^{-1/2}\int \phi_0 \widetilde{\Phi} dP + h_2 m^{-1/2}\int \psi_0 \widetilde{\Psi} dQ + O(h_1^2/n + h_2^2/m) \\ &= h_1 n^{-1/2}\operatorname{Var}_P[\phi_0(X)] + h_2 m^{-1/2}\operatorname{Var}_Q[\psi_0(Y)] + O(h_1^2/n + h_2^2/m).\end{aligned}$$

We deduce that

$$\sqrt{\frac{nm}{m+m}}(\kappa_n(h) - \kappa_n(0)) \longrightarrow \dot\kappa(h) := \langle(\sqrt{1-\rho}, \sqrt{\rho}), (h_1, h_2)\rangle_H,$$

as $n, m \to \infty$ such that $n/(n+m) \to \rho$. It follows that the sequence of parameters $(\kappa_n(h) : h \in H)$ is regular. Furthermore, the adjoint of $\dot\kappa$ is easily seen to be $\dot\kappa^*(b^*) = b^*(\sqrt{1-\rho}, \sqrt{\rho})$, for all $b^* \in \mathbb{R}$, and one has

$$\|\dot\kappa^*(b^*)\|_H^2 = b^*\Big((1-\rho)\operatorname{Var}_P[\phi_0(X)] + \rho \operatorname{Var}_Q[\psi_0(Y)]\Big).$$

The claim now follows from Lemma 91. □

7.C Alternate Proofs of Central Limit Theorems

In this Section, we provide an alternate proof of Theorem 26 which does not rely on our stability bounds in Theorem 18 and Proposition 23. We instead follow the strategy developed by del Barrio and Loubes (2019) for obtaining limit laws of the process $\sqrt{n}(W_2^2(P_n, Q) - W_2^2(P, Q))$. For the sake of brevity, we only prove the one-sample case of Theorem 26(ii), and the remaining assertions of Theorem 26 can be handled similarly. Throughout this section, we abbreviate $\Psi = \Psi^{\mathrm{bc}}$ and $\widehat{P}_n = \widehat{P}_n^{(\mathrm{bc})}$.

We shall make use of the classical Efron-Stein inequality (see for instance Boucheron, Lugosi, and Massart (2013), Theorem 3.1) for bounding the variance of functions of independent random variables, stated as follows.

Lemma 93 (Efron-Stein Inequality). *Let $X_1, X_1', X_2, X_2', \dots, X_n, X_n'$ be independent random variables, and let $R_n = f(X_1, \dots, X_n)$ be a square-integrable function of $X_1, \dots, X_n$. Let*

$$R_{ni}' = f(X_1, \dots, X_{i-1}, X_i', X_{i+1}, \dots, X_n), \quad i = 1, \dots, n.$$

Then,

$$\operatorname{Var}[R_n] \leq \sum_{i=1}^n \mathbb{E}(R_n - R_{ni}')_+^2.$$

With these results in place, we turn to proving the one-sample case of Theorem 26(ii). In view of Lemma 106, it suffices to assume $P \neq Q$, in which case $\operatorname{Var}[\phi_0(X)] > 0$. We abbreviate $\widehat{P}_n = \widehat{P}_n^{(\mathrm{bc})}$, and we begin with the following result.

Proposition 41. Assume the same conditions as Theorem 26(ii). Define

$$R_n = W_2^2(\widehat{P}_n, Q) - \int \phi_0 d\widehat{P}_n.$$

Then, as $n \to \infty$, $n \operatorname{Var}(R_n) \to 0$.

7.C.1 Proof of Proposition 41

Let $X_1' \sim P$ denote a random variable independent of $X_1, \dots, X_n$, and let

$$P_n' = \frac{1}{n}\delta_{X_1'} + \frac{1}{n}\sum_{i=2}^n \delta_{X_i}$$

denote the corresponding empirical measure. Let $\widehat{P}_n'$ be the distribution with density

$$\widehat{p}_n' = \sum_{\zeta\in\Phi} \widehat{\beta}_\zeta' \zeta + \sum_{j=j_0}^{J_n}\sum_{\xi\in\Psi_j} \widehat{\beta}_\xi' \xi = \sum_{j=j_0-1}^{J_n}\sum_{\xi\in\Psi_j} \widehat{\beta}_\xi' \xi, \quad \text{where } \widehat{\beta}_\xi' = \int \xi d\widehat{P}_n',\ \xi \in \Psi,$$

where we write $\Psi_{j_0-1} = \Phi$ for ease of notation. Set

$$R_n' = W_2^2(\widehat{P}_n', Q) - \int \phi_0 d\widehat{P}_n'.$$

By Lemma 93, it will suffice to prove that $n^2\mathbb{E}(R_n - R_n')_+^2 = o(1)$. Let $(\widehat{\phi}_n, \widehat{\psi}_n)$ be a pair of Kantorovich potentials between $\widehat{P}_n$ and Q. Without loss of generality, we may assume that $\int \widehat{\phi}_n d\mathcal{L} = \int \phi_0 d\mathcal{L}$ for all $n \geq 1$. By the Kantorovich duality, we have

$$\begin{aligned} W_2^2(\widehat{P}_n, Q) &= \int \widehat{\phi}_n d\widehat{P}_n + \int \widehat{\psi}_n dQ, \\ W_2^2(\widehat{P}_n', Q) &= \sup_{(\phi,\psi)\in\mathcal{K}} \int \phi d\widehat{P}_n' + \int \psi dQ \\ &\geq \int \widehat{\phi}_n d\widehat{P}_n' + \int \widehat{\psi}_n dQ = W_2^2(\widehat{P}_n, Q) + \int \widehat{\phi}_n d(\widehat{P}_n' - \widehat{P}_n). \end{aligned}$$

It follows that, on the event E_n,

$$R_n - R_n' \leq \int (\widehat{\phi}_n - \phi_0) d(\widehat{P}_n - \widehat{P}_n').$$

In view of Lemma 93, the claim will follow if we are able to show that $n^2\mathbb{E}(R_n - R_n')_+ = o(1)$. Arguing similarly as in the proof of, for instance, Lemma 36, it holds that $\mathbb{P}(\widehat{p}_n = \widetilde{p}_n) \lesssim n^{-3}$. Using this fact and the above inequality, it will suffice to prove that the quantity

$$\Delta_n := n^2\mathbb{E}\left(\int (\widehat{\phi}_n - \phi_0)(\widetilde{p}_n - \widetilde{p}_n')d\mathcal{L}\right)_+^2$$

vanishes as $n \to \infty$. To this end, notice that

$$\int (\widehat{\phi}_n - \phi_0)(\widetilde{p}_n - \widetilde{p}_n')d\mathcal{L} = \int (\widehat{\phi}_n - \phi_0)\left(\sum_{j=j_0-1}^{J_n}\sum_{\xi\in\Psi_j} (\widehat{\beta}_\xi - \widehat{\beta}_\xi')\xi\right)$$

$$= \frac{1}{n} \sum_{j=j_0-1}^{J_n} \sum_{\xi \in \Psi_j} (\xi(X_1) - \xi(X_1')) \int (\widehat{\phi}_n - \phi_0)\xi.$$

Using the locality of the wavelet basis (Lemma 30(ii)) and the Cauchy-Schwarz inequality, we obtain

$$\Delta_n \lesssim J_n \sum_{j=j_0-1}^{J_n} \sum_{\xi \in \Psi_j} \mathbb{E}[\xi^2(X)] \int \big\|\widehat{\phi}_n - \phi_0\big\|^2 |\xi|^2 \lesssim J_n \sum_{j=j_0-1}^{J_n} \sum_{\xi \in \Psi_j} \int \big\|\widehat{\phi}_n - \phi_0\big\|^2 |\xi|^2.$$

In the final step, we again used Lemma 30(ii) together with the fact that p is bounded over $[0,1]^d$ (since $p \in \mathcal{C}^{\alpha-1}([0,1]^d)$), implying that

$$\mathbb{E}[\xi^2(X)] = \int \xi^2(x)p(x)dx \lesssim \int \xi^2(x)dx = 1.$$

By Lemma 101, for all $\xi \in \Psi_j$ and $j \geq j_0$, we have $\mathrm{supp}(\xi) \subseteq I_\xi$ for a rectangle $I_\xi \subseteq [0,1]^d$ satisfying $\mathrm{diam}(I_\xi) \lesssim 2^{-j}$, and $\|\xi\|_{L^\infty(I_\xi)} \lesssim 2^{dj/2}$. Thus,

$$n^2 \mathbb{E}(R_n - R_n')_+^2 \lesssim J_n \sum_{j=j_0-1}^{J_n} 2^{dj} \int_{I_\xi} \big\|\widehat{\phi}_n - \phi_0\big\|^2.$$

Apply the Poincaré inequality in Lemma 38 together with the bound $\mathrm{diam}(I_\xi) \lesssim 2^{-j}$ to deduce

$$n^2 \mathbb{E}(R_n - R_n')_+^2 \lesssim J_n \sum_{j=j_0-1}^{J_n} 2^{(d-2)j} \int_{I_\xi} \big\|\nabla(\widehat{\phi}_n - \phi_0)\big\|^2 \lesssim J_n \sum_{j=j_0-1}^{J_n} 2^{(d-2)j} \big\|\widehat{T}_n - T_0\big\|_{L^2(P)}^2,$$

where the final inequality holds due to the assumption that p has a positive density over $[0,1]^d$, which, due to the continuity of p, implies that there is a constant $\gamma^{-1} > 0$ such that $p \geq \gamma^{-1}$ over $[0,1]^d$. Apply Theorem 19 to deduce that

$$n^2 \mathbb{E}(R_n - R_n')_+^2 \lesssim J_n \sum_{j=j_0-1}^{J_n} 2^{(d-2)j} \big\|\widehat{T}_n - T_0\big\|_{L^2(P)}^2 \lesssim J_n \left(2^{J_n(d-2-2\alpha)} \vee \frac{(\log n)^2}{n}\right).$$

Since $d < 2(\alpha+1)$ and $J_n \asymp \log n$, the above display is of order $o(1)$, thus the claim follows from Lemma 93. □

To prove the claim from here, write

$$\sqrt{n}\Big(W_2^2(\widehat{P}_n, Q) - \mathbb{E}W_2^2(\widehat{P}_n, Q)\Big) = \sqrt{n} \int \phi_0(\widehat{p}_n - p_{J_n}) + \sqrt{n}\Big(R_n - \mathbb{E}[R_n]\Big),$$

where recall that $p_{J_n} = \mathbb{E}[\widehat{p}_n]$. It follows from Proposition 41 that the final term of the above display converges to zero in probability. Furthermore, $\sqrt{n}\int \phi_0(\widehat{p}_n - p_{J_n}) \rightsquigarrow N(0, \mathrm{Var}[\phi_0(X)])$ by Lemma 89. Combining these facts with the bias bound of Theorem 19, the claim follows. □

Part III

Optimal Transport in Action

Chapter 8

An Application to the Search for Pairs of Higgs Bosons

Our aim in this final chapter is to provide an application of the methods developed in this thesis to a statistical problem arising in high energy physics, and more specifically in the search for new physical phenomena in collider experiments.

8.1 Introduction

The Standard Model (SM) of high energy physics is a theory describing the interactions between elementary particles—the building blocks of matter. One key component of the SM is the presumed existence of a quantum field responsible for generating mass in certain elementary particles. This field is known as the Higgs field, originally theorized by Higgs (1964), Englert and Brout (1964). Excitations of the Higgs field produce particles, known as Higgs bosons, which were the subject of an intensive search by experimental particle physicists ever since the mid 1970s. In July 2012, two independent experiments at the Large Hadron Collider (LHC) at CERN (the European Organization for Nuclear Research) announced the observation of a new particle consistent with the SM Higgs boson (ATLAS, 2012; CMS, 2012). Having discovered this Higgs-like particle, current work is concerned with detailed studies of its properties, in order to confirm or refute those predicted by the SM. One such property is the so-called Higgs boson *self-coupling*, whereby a single excitation of the Higgs field can split into two Higgs bosons without intermediate interactions with other particles. Observing this phenomenon would provide compelling new information regarding the mechanism of particle mass generation. This chapter is concerned with some of the statistical challenges posed by its search.

The LHC is housed in a massive underground tunnel in which two counter-rotating beams of protons are accelerated to nearly the speed of light. When these protons collide, new particles are formed, and their paths within particle detectors are recorded. Individual collisions are referred to as *events*. An event in which two Higgs bosons are generated is called a *double Higgs* *(or di-Higgs) event*. The Higgs boson is a highly unstable particle; whenever it is produced, it decays into other particles almost immediately, making di-Higgs production impossible to observe directly.

The Higgs boson most commonly decays into a pair of so-called bottom quarks (b-quarks). An event in which four b-quarks are observed is thus a candidate di-Higgs event, but could also have arisen from various other physical processes that produce four b-quarks. We say that a di-Higgs event in which the Higgs bosons decay into four b-quarks is a *signal event*, while any other event tagged as having four b-quarks is called a *background event*. The problem of searching for double Higgs boson production reduces to testing whether the proportion of signal events is nonzero among the observed data. As we describe in Section 8.3, carrying out this test is a well-understood statistical task when the distributions of both background and signal events are known. While the di-Higgs signal distribution can be approximated to sufficient accuracy with first-principles simulation, simulating the background distribution suffers from large high-order corrections which are computationally intractable (Di Micco et al., 2020). Instead, the background distribution must be estimated using observed data. This is known as the problem of *data-driven background modeling*, which is the main subject of this chapter.

As stated, the background distribution is not a statistically identifiable quantity without further assumptions, due to the potential presence of an unknown proportion of signal in the data. Any analysis strategy must therefore make some modeling assumptions to make the background estimation problem tractable. As we discuss below, it is standard to assume that the background distribution is related in some way to the distribution of certain *auxiliary events*, which in turn is identifiable. An example of useful auxiliary events is those consisting of less than four observed b-quarks, since they are unlikely to be signal events, but are kinematically similar to the background events of interest (Bryant, 2018). Stated differently, the distribution of auxiliary events is an identifiable estimand which has undergone a *distributional shift* relative to the non-identifiable background distribution of interest. If the analyst has access to a sample of auxiliary events, its empirical distribution provides a first naive approximation of the desired background distribution. To obtain a more precise estimate, one must correct for the distributional shift.

As we discuss in Section 8.1.2, the most widely-used method for correcting this distributional shift is based on an estimate of the *density ratio* between the background and auxiliary events. This method typically first estimates the density ratio in a signal-free region of the phase space, known as the *Control Region*, and then extrapolates it to the region of primary interest, known as the *Signal Region*. Any deviation of this extrapolated density ratio from unity is used to correct the distributional shift undergone by the auxiliary sample. This extrapolation can be viewed as an instance of transfer learning (Weiss, Khoshgoftaar, and Wang, 2016). While a careful choice of the density ratio estimator can greatly improve the accuracy of this extrapolation, it clearly cannot lead to a consistent estimator if the distribution in the Signal Region is unconstrained relative to its counterpart in the Control Region. This procedure thus places an implicit modeling assumption on the underlying distributions, which is challenging to quantify and verify in practice. Nevertheless, variants of this procedure have been used in each of the most recent di-Higgs searches in the four b-quark final state (e.g. ATLAS (2018), ATLAS (2019), ATLAS (2021), CMS (2022), ATLAS (2022)). This raises the important need for cross-checking the modeling assumption made by such an approach.

8.1.1 Our Contributions

This chapter develops a new methodology for data-driven background modeling in di-Higgs boson searches. Our approach is fully nonparametric, and does not involve the extrapolation of density ratios. It hinges upon a characteristic modeling assumption, which is complementary to that of the density ratio method. These two distinct methods can thus serve as cross-checks for each other in di-Higgs searches, an important benefit that will increase the analyst's trust in the obtained background estimates.

Our approach is based on the optimal transport problem (Villani, 2003) between multidimensional distributions of collider events. Optimal transport has already proven to be a powerful tool for transfer learning in classification problems (Courty et al., 2016), and here we propose to use it rather differently to correct distributional shifts between estimates of the auxiliary and background distributions. Our method involves out-of-sample estimation of optimal transport maps, for which we consider two different estimators. The first is based on nearest neighbor extrapolation, and was already introduced in Chapter 5. Our second estimator appears to be new, and leverages some strengths of the density ratio approach.

Unlike our earlier work on optimal transport map estimation, however, we will find it useful in this chapter to work with a non-quadratic cost function. We will instead take the cost to be a metric over the space of collider events which was proposed by Komiske, Metodiev, and Thaler (2019). This metric is itself obtained through the optimal transport problem of matching clusters of energy deposits in collision events. Our approach therefore involves a nested use of optimal transport.

We illustrate the empirical performance of our method, as compared to the density ratio approach, on realistic simulated collider data. We observe that both approaches lead to quantitatively similar background estimates, despite the complementarity of their underlying modeling assumptions. In particular, this study illustrates how our methods can be used to cross-check each other in practice.

8.1.2 Related Work

Di-Higgs boson production has been the subject of numerous recent searches by the ATLAS and CMS collaborations at the LHC—we refer to the recent survey paper of Di Micco et al. (2020) for an overview. The four b-quark final state is the most common decay channel for di-Higgs events, but suffers from a large multijet background. As described previously, each of the most recent searches in this final state performed data-driven background estimation by first estimating a density ratio in a Control Region, and extrapolating it to the Signal Region. Certain searches, such as ATLAS (2019), estimate the density ratio using heuristic one-dimensional reweighting schemes, while others, such as CMS (2022), use off-the-shelf multivariate classifiers for this purpose. More broadly, the ihe idea of estimating density ratios using classifiers has a long history in statistics—see for instance Fix and Hodges (1951), Silverman and Jones (1989), Qin (1998), Cheng and Chu (2004), Kpotufe (2017)—and appears in a variety of applications in experimental particle physics (e.g., Cranmer, Pavez, and Louppe (2015), Brehmer et al. (2020), CMS (2022)). Classification-based estimators have the practical advantage of circumventing the need for high-dimensional density estimation, which can be particularly challenging to perform over the space of collider events.

Rather than density ratios, the key object of interest in our new methodology is a nested optimal transport problem. Hierarchical optimal transport problems of this kind have recently been used for other tasks, such as multilevel clustering (Ho et al., 2017, 2019; Huynh et al., 2021) and multimodal distribution alignment (Lee et al., 2019). Very recently, optimal transport has also been used in high energy physics for calibrating stochastic simulators (Pollard and Windischhofer, 2022), for purposes of exploratory data analysis (Komiske, Metodiev, and Thaler, 2019; Komiske et al., 2020; Cai et al., 2020), and for the purpose of defining a geometry on the space of collider events (Komiske, Metodiev, and Thaler, 2020). We also note that optimal transport has implicitly been used for one-dimensional template morphing in the early work of Read (1999).

Beyond the search of di-Higgs boson production, we emphasize that the question of data-driven background estimation arises in a variety of problems in experimental high-energy physics, where our methodologies could also potentially be applied. We refer to the book Behnke et al. (2013) for a pedagogical introduction to statistical aspects of the subject; see also Appendix 1 of Lyons (1986). Finally, we mention some recent advances on the widely-used sPlot (Barlow, 1987; Pivk and Le Diberder, 2005; Borisyak and Kazeev, 2019; Dembinski et al., 2022) and ABCD (Alison, 2015; ATLAS, 2015; Choi and Oh, 2021; Kasieczka et al., 2021) techniques for background estimation, the latter of which can be viewed as a precursor to the methods developed in this chapter.

8.2 Background

8.2.1 LHC Experiments and di-Higgs Boson Production

The LHC is the largest particle collider in the world, consisting of a 27 kilometer-long tunnel in which two counter-rotating beams of protons are accelerated to nearly the speed of light. These particles are primarily collided in one of four underground detectors, named ALICE, ATLAS, CMS and LHCb. ATLAS and CMS are general-purpose detectors used for a wide range of physics analyses, including Higgs boson-related searches, while ALICE and LHCb focus on specific physics phenomena. We focus on the CMS detector in what follows, but similar descriptions can be made for the ATLAS detector.

When two protons collide, their energy is converted into matter, in the form of new particles. The goal of the CMS (Compact Muon Solenoid) detector is to measure the momenta, energies and types of such particles. To measure their momenta, CMS is built around a giant superconducting solenoid magnet, depicted in Figure 8.1, which deforms the trajectories of particles as they move from the center of the detector outward through a silicon tracker. The

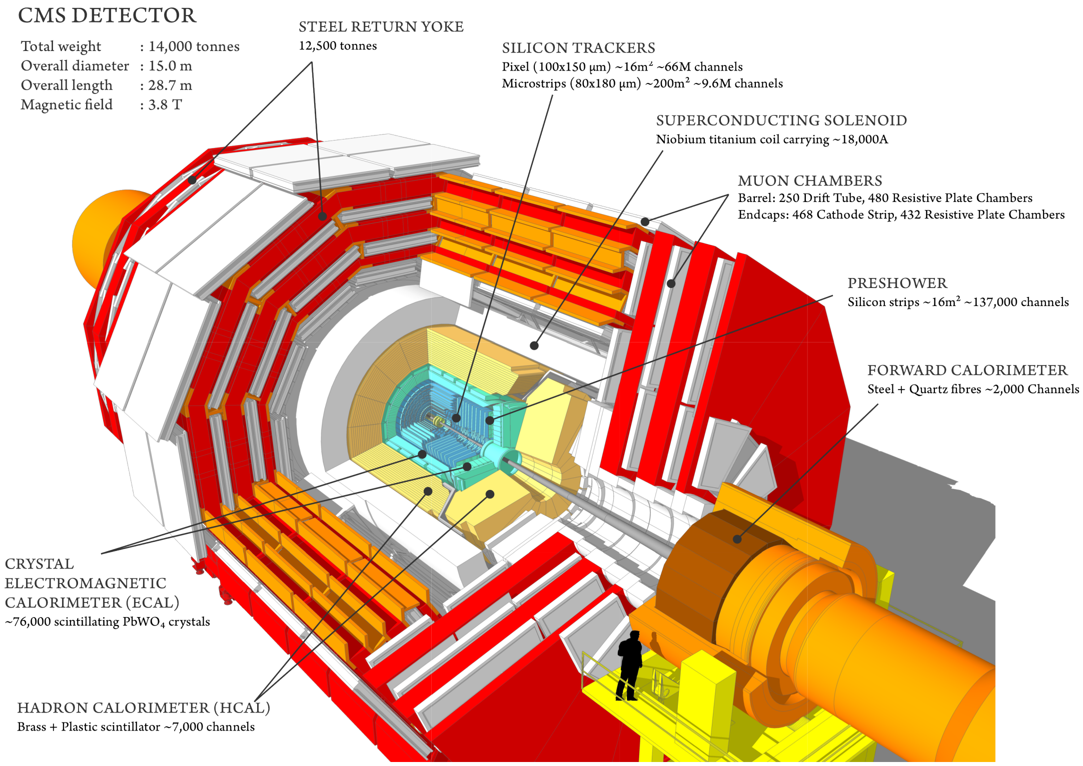
CMS DETECTOR
Total weight : 14,000 tonnes
Overall diameter : 15.0 m
Overall length : 28.7 m
Magnetic field : 3.8 T
STEEL RETURN YOKE
12,500 tonnes
SILICON TRACKERS
Pixel (100x150 μm) ~16m² ~66M channels
Microstrips (80x180 μm) ~200m² ~9.6M channels
SUPERCONDUCTING SOLENOID
Niobium titanium coil carrying ~18,000A
MUON CHAMBERS
Barrel: 250 Drift Tube, 480 Resistive Plate Chambers
Endcaps: 468 Cathode Strip, 432 Resistive Plate Chambers
PRESHOWER
Silicon strips ~16m² ~137,000 channels
FORWARD CALORIMETER
Steel + Quartz fibres ~2,000 Channels
CRYSTAL ELECTROMAGNETIC CALORIMETER (ECAL)
~76,000 scintillating PbWO₄ crystals
HADRON CALORIMETER (HCAL)
Brass + Plastic scintillator ~7,000 channels

Figure 8.1: Illustration of the CMS detector (Sakuma and McCauley, 2014). Counter-rotating beams of protons are made to collide in the center of the detector. The trajectory and mass of each particle emanating from the collision is then recorded.

extent to which the trajectory of a charged particle is bent depends on its momentum and can hence be used to measure the momentum. After the silicon tracker, CMS consists of several layers of calorimeters which measure the energies of the particles. We refer to CMS (2008) for a complete description of the CMS detector.

Proton-proton collisions give rise to highly unstable particles which decay almost instantly into more stable particles. The detector is only able to observe these longer-lived particles. By measuring their energies and momenta, insight can be gained into the physical properties of the unstable particles from which they originate.

The Higgs boson is an example of an unstable particle, which decays within approximately 10^{-22} seconds. The SM predicts that a Higgs boson decays into a pair of bottom quarks (b-quarks) 60% of the time, and this decay channel has indeed been observed experimentally (ATLAS, 2018; CMS, 2018). Other channels which have been observed experimentally include the decay of a Higgs boson into pairs of photons (ATLAS, 2018; CMS, 2018), W bosons (ATLAS, 2018; CMS, 2019), Z bosons (ATLAS, 2018; CMS, 2018), and tau leptons (ATLAS, 2019; CMS, 2018). The SM further predicts the rare possibility that two Higgs bosons can be produced simultaneously, and this chapter is concerned with the statistical challenges arising in the search for this process, which has yet to be observed experimentally. If this process were to occur, the two resulting Higgs bosons would each, in turn, be most likely to decay into two b-quarks, thus making four b-quarks the most common final state of di-Higgs boson events. We focus on this decay channel (abbreviated HH$\to$ $4b$) throughout this chapter. We note that b-quarks form into bound states with other quarks called b-hadrons which are themselves unstable, and rapidly decay into collimated sprays of stable particles called b-jets, which can be efficiently identified by the CMS detector (CMS, 2018).

8.2.2 Collider Events and the CMS Coordinate System

Particles measured by the CMS detector are typically represented in spherical coordinates. Given a particle with momentum vector $p = (x, y, z) \in \mathbb{R}^3$, its azimuthal angle $\phi \in [0, 2\pi)$ is defined as the angle increasing from the positive x-axis to the positive y-axis, while the polar angle $\theta \in [0, \pi)$ is increasing from the positive z-axis to the positive y-axis. The length of its projection onto the (x, y) plane is called the *transverse momentum* p_T. It is common to replace the polar angle θ by the *pseudorapidity* of the particle, given by $\eta = -\log(\tan(\theta/2))$.

In addition to the variables p_T, η and ϕ, the rest mass m of each particle can be obtained from the energy measurements made by the calorimeters in the CMS detector. Altogether, a particle jet is analyzed as a single point in this coordinate system, and encoded as a four-dimensional vector (p_T, η, ϕ, m). In our search channel, collisions lead to multiple, say $K \geq 1$, jets measured by the detector, which may be encoded as the $4K$-dimensional vector $(p_{T_i}, \eta_i, \phi_i, m_i : 1 \leq i \leq K)$. We opt for an alternative notation, which will be particularly fruitful for the purpose of defining a metric between collider events in Section 8.5.2. Specifically, an event will henceforth be represented by the discrete measure

$$g = \sum_{i=1}^{K} p_{T_i} \delta_{(\eta_i, \phi_i, m_i)}, \tag{8.1}$$

where δ_x denotes the Dirac measure placing unit mass at a point $x \in \mathbb{R}^3$. In particular, the representation (8.1) emphasizes the invariance of an event with respect to the ordering of its jets. The transverse momenta p_{T_i} may be viewed as a proxy for the energy of each jet, thus the total measure of g denotes its total energy, denoted $s_T = \sum_{i=1}^{K} p_{T_i}$. The set of events with K jets of interest is denoted by

$$\mathcal{G}^{(K)} = \left\{ \sum_{j=1}^{K} p_{T_j} \delta_{(\eta_j, \phi_j, m_j)} : p_{T_j}, m_j > 0,\ \phi_j, \eta_j \in \mathbb{R},\ 1 \leq j \leq K \right\},$$

where the definition of ϕ_j is extended from $[0, 2\pi)$ to the entire real line by 2π-periodicity. In the context of double Higgs boson production in the four b-jet final state, the choice $K = 4$ will be most frequently used, and in this case we simply write $\mathcal{G} = \mathcal{G}^{(4)}$.

Finally, we note that events are deemed invariant under the orientation of the x- and z-axes. This fact, together with the periodicity of the angle ϕ, implies that two events $g =$

$\sum_{j=1}^K p_{T_j}\delta_{(\eta_j,\phi_j,m_j)} \in \mathcal{G}^{(K)}$ and $g' = \sum_{j=1}^K p'_{T_j}\delta_{(\eta'_j,\phi'_j,m'_j)} \in \mathcal{G}^{(K)}$ may be deemed equivalent if they are mirror-symmetric in η, ϕ, as well as rotationally symmetric in ϕ, that is, if there exist $\Delta \in 2\pi\mathbb{Z}$ and $\iota_1, \iota_2 \in \{-1, 1\}$ such that

$$\sum_{j=1}^K p_{T_j}\delta_{(\iota_1\eta_j,\Delta+\iota_2\phi_j,m_j)} = \sum_{j=1}^K p'_{T_j}\delta_{(\eta'_j,\phi'_j,m'_j)}. \tag{8.2}$$

Formally, we define an equivalence relation $\simeq$ between events in $\mathcal{G}^{(K)}$, such that $g \simeq g'$ if and only if there exist Δ, ι_1, ι_2 for which (8.2) holds.

8.3 Problem Formulation

8.3.1 Overview of Signal Searches at the LHC

In order to make inferences about the presence or absence of a signal process in collider data, event counts are commonly analyzed as binned Poisson point processes. While we focus on the setting of double Higgs boson production in the four b-quark final state, the description that follows is representative of a wide range of signal searches for high-energy physics experiments.

Let ν_0 denote a σ-finite Borel measure over the state space $\mathcal{G}$ of collider events, with respect to a fixed choice of Borel σ-algebra on $\mathcal{G}$ denoted $\mathbb{B}(\mathcal{G})$. Let F denote an inhomogeneous Poisson point process (Reiss, 2012) with a nonnegative intensity function $f \in L^2(\mathcal{G})$ on $\mathcal{G}$, that is, F is a random point measure on $\mathcal{G}$ such that

1. $F(A) \sim \text{Poisson}(\lambda(A))$, where λ is the intensity measure induced by f, defined by $\lambda(A) = \int_A f d\nu_0$ for all $A \in \mathbb{B}(\mathcal{G})$;
2. $F(A_1), \dots, F(A_\ell)$ are independent for all pairwise disjoint sets $A_1, \dots, A_\ell \in \mathbb{B}(\mathcal{G})$, for all integers $\ell \geq 1$.

Every four b-jet collision event is either a *signal event*, namely an event arising from two Higgs bosons, or a *background event*, arising from some other physical process. Letting $\mu \geq 0$ denote the rate of signal events, we write the intensity measure λ as

$$\lambda(\cdot) = \beta_4(\cdot) + \mu\sigma(\cdot),$$

where β_4 and σ, respectively, denote nonnegative background and signal intensity measures. σ is typically normalized such that the value $\mu = 1$ corresponds to the theoretical prediction of the signal rate. The measures β_4 and σ typically depend on nuisance parameters related to the calibration of the detector, the uncertain parameters of certain physical processes, such as the parton distribution functions of the proton (Placakyte, 2011), and so on. We suppress the dependence on such nuisance parameters for ease of exposition. The parameter μ is of primary interest, since non-zero values of μ indicate the existence of signal events. A search for the signal process therefore reduces to testing the following hypotheses on the basis of observations from the Poisson point process F:

$$H_0 : \mu = 0 \quad \text{vs.} \quad H_1 : \mu > 0. \tag{8.3}$$

Given a sequence $G_1, G_2, \dots$ of observed events, we may write $F = \sum_{i=1}^{M} \delta_{G_i}$, where $M \sim \text{Poisson}(\lambda(\mathcal{G}))$ is independent of the observations, which satisfy

$$G_1, G_2, \dots \overset{\text{iid}}{\sim} \lambda/\lambda(\mathcal{G}) = \epsilon S + (1-\epsilon)P_4. \tag{8.4}$$

Here, $S = \sigma/\sigma(\mathcal{G})$ and $P_4 = \beta_4/\beta_4(\mathcal{G})$ denote the respective signal and background distributions, and $\epsilon = \mu\sigma(\mathcal{G})/\lambda(\mathcal{G})$ the proportion of signal events.

The Poisson point process F is often binned in practice. Let $\xi : \mathcal{G} \to \mathcal{A} \subseteq \mathbb{R}$ denote a dimensionality reduction map, to be discussed below, which will be used to bin the point process using univariate bins. Let $\{I_j\}_{j=1}^{J}$ denote a collection of bins forming a partition of $\mathcal{A}$, and define the event counts

$$D_j = F\big(\xi^{-1}(I_j)\big) = \big|\{1 \leq i \leq M : \xi(G_i) \in I_j\}\big|, \quad j = 1, \dots, J. \tag{8.5}$$

The definition of F implies that the random variables D_j are independent and satisfy

$$D_j \sim \text{Poisson}\big(B_j + \mu S_j\big), \quad j = 1, \dots, J, \tag{8.6}$$

where $B_j = \beta_4(\xi^{-1}(I_j))$ and $S_j = \sigma(\xi^{-1}(I_j))$.

The likelihood ratio test with respect to the joint distribution of $D_1, \dots, D_J$ is typically used to test the hypotheses (8.3) (ATLAS, CMS, and Higgs Combination Group, 2011). The binned likelihood function for the parameter μ is given by

$$L(\mu) = \prod_{j=1}^{J} \frac{\big(B_j + \mu S_j\big)^{D_j}}{D_j!} e^{-\left(B_j + \mu S_j\right)}. \tag{8.7}$$

Di-Higgs events are rare in comparison to background events, thus the signal-to-background ratio is low. At the time of writing, values of M which are typically observed at the LHC may be too small for any test to have power in rejecting the null hypothesis in (8.3) at desired significance levels (Di Micco et al., 2020). Analyses which fail to reject H_0 instead culminate in an upper confidence bound on μ, also known as an upper limit (ATLAS, CMS, and Higgs Combination Group, 2011).

The power of the likelihood ratio test for (8.3) may be increased by choosing a function ξ which maximizes the separation between background and signal event counts across the J bins. Informally, the optimal such choice of ξ is given by

$$\xi(g) = \mathbb{P}(G \text{ is a Signal Event} | G = g), \tag{8.8}$$

which may be estimated using a multivariate classifier, such as a neural network or boosted decision trees, for discriminating background events from signal events.

The signal intensity measure σ is theoretically predicted by the SM, and can be approximated well using Monte Carlo event generators (Di Micco et al., 2020). The background intensity β_4 is, however, intractable due to the strongly interacting nature of quantum chromodynamics (QCD) in which events with the four b-quark final state can be produced via an enormous number of relevant and complex pathways. The intensity measure β_4, or its binned analogue $(B_j)_{j=1}^{J}$, must therefore be estimated from the collider data itself, which we refer to as *data-driven background modeling*. This problem is the primary focus of this chapter.

8.3.2 Setup for Data-Driven Background Modeling

The aim of this chapter is to develop data-driven estimators of the background intensity measure β_4. The primary challenge is the fact that the sample $G_1, \dots, G_M$ is contaminated with an unknown proportion ϵ of signal events. The background estimation problem is thus statistically unidentifiable as stated, and it will be necessary to impose further modeling assumptions.

In order to formulate these assumptions and our resulting background modeling methods, we assume that the analyst has access to a second Poisson Point Process $T = \sum_{i=1}^N \delta_{H_i}$ consisting of auxiliary events which were tagged by the CMS detector as having four jets, of which exactly three are b-jets. We refer to such events as "$3b$ events", as opposed to "$4b$ events" which were identified as having four b-jets[1]. We stress that the terms $3b$ and $4b$ do not refer to the true number of b-quarks arising from the collision, rather the number of b-jets identified by the detector. As we discuss in Section 8.6, the majority of $3b$ events in fact arise from the hadronization of two b-quarks and two charm or light quarks, while a small proportion arise from four b-quarks[2]. For the purpose of a discovery analysis, the $3b$ sample $H_1, \dots, H_N$ can therefore be treated as having a negligible proportion of signal events (Bryant, 2018; CMS, 2022). We treat this proportion as zero for sake of exposition. We henceforth denote the intensity measure of the point process T by β_3, and we denote by $P_3 = \beta_3/\beta_3(\mathcal{G})$ the corresponding probability distribution of the observations $H_1, H_2, \dots$.

The kinematics of $3b$ events are similar, but not equal, to those of $4b$ background events (CMS, 2022). Unlike β_4, however, the intensity measure β_3 is an identifiable estimand due to the lack of signal events in the point process T. Any consistent estimator $\widehat{\beta}_3$ of β_3 can be used to provide a zeroth-order approximation of β_4 (up to a correction for normalization). This approximation is, however, insufficiently accurate to be used as a final estimate of β_4 and our goal is to develop statistical methods for correcting this naive background estimate.

Recall that the four b-jets of any signal event $g \in \mathcal{G}$ are naturally paired, with each pair arising from a Higgs boson. The true pairing of the jets is unknown to the detector; however, it may be approximated, for instance using an algorithm due to Bryant (2018). We use the same pairing algorithm in our work. Given as input an event g, this deterministic algorithm outputs one among the three distinct unordered pairs of measures $\{g^1, g^2\} \subseteq \mathcal{G}^{(2)}$ which satisfy $g = g^1 + g^2$. We refer to g^1 and g^2 as *dijets*. When g is a signal event, we expect that each dijet arose from a decay of a Higgs boson, whereas when g is a background event, we expect that at least one of the two dijets arose from the decay of a different particle.

The Higgs boson is known to have mass m_H approximately equal to 125 GeV (ATLAS, 2012; CMS, 2012). It follows that the two dijets should approximately satisfy $m(g^1) \approx m(g^2) \approx m_H$, where $m(a)$ denotes the invariant mass[3] of any $a \in \mathcal{G}^{(K)}$, $K \geq 1$. Large deviations of the dijet

[1] $3b$ events were used for background estimation in the HH→ $4b$ channel in the recent analysis of CMS (2022). "$2b$ events" consisting of two, rather then three, b-tagged jets have been used in other recent analyses (e.g. ATLAS (2019, 2021)), and our description also applies to such events with only formal changes.

[2] As a result, the expected rate of production of $3b$ events $\mathbb{E}[N]$ is typically higher than that of $4b$ events $\mathbb{E}[M]$ by an order of magnitude; cf. Section 8.6

[3] If E denotes the sum of the energies of the constituent jets of a, and p denotes the magnitude of the sum of

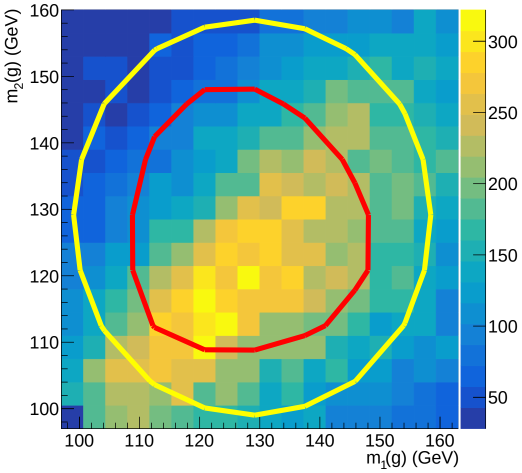

Figure 8.2: Illustration of the Control and Signal Regions. The two-dimensional histogram represents simulated 4b collider events described in Section 8.6, plotted in terms of their dijet invariant masses. We emphasize that this is a low-dimensional representation; the events considered in this work are 16-dimensional. The red line indicates the boundary of the Signal Region, while the annulus bounded by the yellow and red lines represents the Control Region. The constants σ_c, r_c and κ_s used in this figure are stated in Section 8.6.

invariant masses from 125 GeV indicate that g is not a signal event. This provides a heuristic for determining events among $G_1, \dots, G_M$ which are unlikely to be signal events. To elaborate, we form subsets $\mathcal{G}_c, \mathcal{G}_s \subseteq \mathcal{G}$ such that $\mathcal{G}_c \cap \mathcal{G}_s = \emptyset$, where $\mathcal{G}_s$ is called the *Signal Region*, containing events with dijet masses near m_H, and $\mathcal{G}_c$ is called the *Control Region*, containing all other events which will be used in the analysis. We follow Bryant (2018) and employ the following specific definitions of $\mathcal{G}_c$ and $\mathcal{G}_s$:

$$\mathcal{G}_s = \left\{ g \in \mathcal{G} : \sqrt{\left(1 - \frac{m_H}{m(g^1)}\right)^2 + \left(1 - \frac{m_H}{m(g^2)}\right)^2} \leq \kappa_s \right\}, \tag{8.9}$$

$$\mathcal{G}_c = \left\{ g \in \mathcal{G} : \sqrt{\Big(m(g^1) - \sigma_c m_H\Big)^2 + \Big(m(g^2) - \sigma_C m_H\Big)^2} \leq r_c \right\} \setminus \mathcal{G}_s, \tag{8.10}$$

for some constants $\sigma_c, r_c, \kappa_s > 0$. These regions are illustrated in Figure 8.2. We similarly partition the Poisson intensity measures β_3, β_4, by defining for all $A \in \mathbb{B}(\mathcal{G})$,

$$\beta_j^c(A) = \beta_j(A \cap \mathcal{G}_c), \quad \beta_j^s(A) = \beta_j(A \cap \mathcal{G}_s), \quad j = 3, 4.$$

These four measures are illustrated in Figure 8.3. Furthermore, we assume for ease of exposition

their momentum vectors, then the invariant mass of e is defined by $m(a) = \sqrt{E^2 - p^2}$ (Hagedorn, 1964).

that these measures are absolutely continuous with respect to the dominating measure ν_0, and we let $b_j^a = d\beta_j^a/d\nu_0$ for all $j = 3, 4$ and $a = c, s$.

Recall that the collider events associated with the intensity measures β_3^c and β_3^s are signal-free by construction, and those from β_4^c are also signal-free under the assumption that negligibly few signal events will fall outside of $\mathcal{G}_s$. These three intensity measures can therefore be estimated directly by means of their empirical intensity functions. We have thus reduced the background modeling problem to that of estimating β_4^s, given estimates of β_3^c, β_3^s and β_4^c.

To this end, we will partition the samples into the sets

$$\begin{aligned} \{G_1^s, \dots, G_{m_s}^s\} &:= \{G_1, \dots, G_M\} \cap \mathcal{G}_s, \qquad & \{H_1^s, \dots, H_{n_s}^s\} &:= \{H_1, \dots, H_N\} \cap \mathcal{G}_s, \\ \{G_1^c, \dots, G_{m_c}^c\} &:= \{G_1, \dots, G_M\} \cap \mathcal{G}_c, \qquad & \{H_1^c, \dots, H_{n_c}^c\} &:= \{H_1, \dots, H_N\} \cap \mathcal{G}_c, \end{aligned}$$

where $M = m_c + m_s$ and $N = n_c + n_s$. Furthermore, let

$$\beta_{3,n_c}^c = T|_{\mathcal{G}_c} = \sum_{i=1}^{n_c} \delta_{H_i^c}, \quad \beta_{3,n_s}^s = T|_{\mathcal{G}_s} = \sum_{i=1}^{n_s} \delta_{H_i^s}, \quad \beta_{4,m_c}^c = F|_{\mathcal{G}_c} = \sum_{i=1}^{m_c} \delta_{G_i^c}$$

denote the empirical estimators of the measures $\beta_3^c, \beta_3^s, \beta_4^c$, illustrated in the background of Figure 8.3. As previously noted, the measure β_3^s provides a zeroth-order approximation of β_4^s (after a normalization correction), thus a naive first estimate of β_4^s is given by β_{3,n_s}^s. As we shall see in the simulation study of Section 8.6, this approximation is insufficiently accurate to be used as a final estimator. Our methodologies improve upon it by modeling the discrepancy between the $3b$ and $4b$ distributions in the Control Region via $\beta_{4,m_c}^c, \beta_{3,n_c}^c$, and then using that information in the Signal Region to improve the accuracy of β_{3,n_s}^s as an estimator of β_4^s.

Once we are able to derive an estimator $\widehat{\beta}_4^s$ of β_4^s, based on the signal-free observations $G_1^c, \dots, G_{m_c}^c$, $H_1^s, \dots, H_{n_s}$, $H_1^c, \dots, H_{n_c}^c$, we may define the fitted histogram $\widehat{B}_j = \widehat{\beta}_4^s(\xi^{-1}(I_j))$, $j = 1, \dots, J$. One may then test the hypotheses (8.3) using the likelihood ratio test, based on the following modification of the likelihood function in equation (8.7),

$$\widetilde{L}(\mu) = \prod_{J=1}^{J} \frac{\left(\widehat{B}_j + \mu S_j\right)^{D_j^s}}{D_j^s!} e^{-\left(\widehat{B}_j + \mu S_j\right)}, \quad \text{where } D_j^s = \left|\{1 \leq i \leq m_s : \xi(G_i^s) \in I_j\}\right|. \tag{8.11}$$

Here, $\widetilde{L}$ can be viewed as a restriction of the likelihood L to the Signal Region. Notice that $\widehat{B}_j$ is independent of D_k^s, for any j, k. In practice, it is also necessary to incorporate statistical and systematic uncertainties pertaining to the estimator $\widehat{B}_j$ into the hypothesis testing procedure (ATLAS, CMS, and Higgs Combination Group, 2011). Since formal uncertainty quantification for background modeling is beyond the scope of this work, we omit further details, and provide further discussion of this point in Section 8.7.

The primary difficulty remaining in the testing problem (8.3) is that of deriving estimators of the background intensity measure β_4^s. In what follows, we describe two classes of estimators for β_4^s: one based on density ratio estimation (Section 8.4), and the second based on optimal

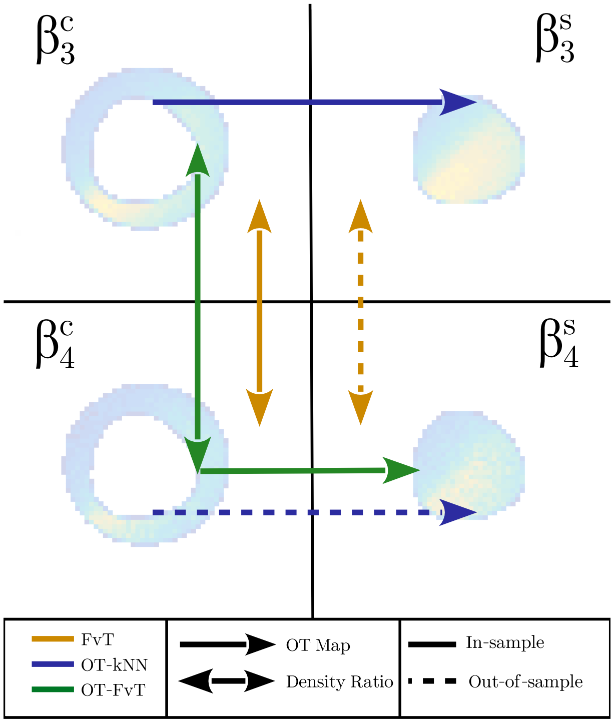

Figure 8.3: Illustration of the four Poisson intensity measures $\beta_3^c, \beta_3^s, \beta_4^c, \beta_4^s$, among which only the latter is nontrivial to estimate, and summary of the three methods developed in this chapter for estimating β_4^s. The method FvT (Four vs. Three) estimates the ratio of the two densities in the Control Region using a classifier, and then extrapolates it into the Signal Region using out-of-sample evaluations of the classifier. The OT-kNN (Optimal Transport–k Nearest Neighbors) method produces an estimator $\widehat{T}$ of the optimal transport map T between the $3b$ Control and Signal Region distributions, and evaluates this estimator out-of-sample on an estimator of the 4b Control Region distribution. The out-of-sample evaluation of $\widehat{T}$ is performed using nearest-neighbor extrapolation. The OT-FvT (Optimal Transport–Four vs. Three) method combines these ideas: first, it uses the classifier to produce an estimator of β_4^c with the same support as β_{3,n_c}^c, and second, it pushes forward this estimator through $\widehat{T}$, thereby avoiding out-of-sample evaluations of both the classifier and optimal transport map. The background of the figure consists of bivariate histograms of simulated $3b$ and $4b$ samples in the Control and Signal Regions, plotted in terms of their dijet invariant masses, as in Figure 8.2.

transport (Section 8.5). The former is the most common approach to di-Higgs background modeling, and will later be referred to as the FvT method. The latter is new, and we will discuss two different instances of this approach, which will later be referred to as the OT-kNN and OT-FvT methods. These three distinct estimators are summarized in Figure 8.3.

8.4 Background Modeling via Density Ratio Extrapolation

The discrepancy between $3b$ and $4b$ background distributions may be directly quantified in the Control Region, where no signal events are present. Under a suitable modeling assumption, this discrepancy may be extrapolated into the Signal Region to produce a correction of the $3b$ signal region intensity measure β_3^s, leading to an estimate of β_4^s. This general strategy forms the basis of most background modeling methodologies used in recent di-Higgs searches, as discussed in Section 8.1. in this brief section, we recall how this approach may be carried out using a classifier for discriminating $3b$ and $4b$ events.

Let E denote a random collider event, arising from either the $3b$ or $4b$ distributions, and define the latent binary random variable Z indicating the component membership of E. More specifically, let Z be a Bernoulli random variable with success probability $\mathbb{P}(Z = 1) = \beta_4(\mathcal{G})/(\beta_4(\mathcal{G}) + \beta_3(\mathcal{G}))$, and let E be generated according to the mixture model

$$E|Z = 0 \sim P_3, \quad E|Z = 1 \sim P_4.$$

Setting $\psi(g) = \mathbb{P}(Z = 1|E = g)$ for all $g \in \mathcal{G}$, it follows from Bayes' Rule that

$$\frac{b_4^c(g)}{b_3^c(g)} = \frac{\psi(g)}{1 - \psi(g)}, \quad g \in \mathcal{G}_c, \tag{8.12}$$

where we recall that b_j^c denotes the intensity function associated to β_j^c, $j = 3, 4$. Therefore,

$$\beta_4^c(A) = \int_A \frac{\psi(g)}{1 - \psi(g)} d\beta_3^c(g), \quad A \in \mathbb{B}(\mathcal{G}_c). \tag{8.13}$$

Equations (8.12–8.13) are a reformulation for our context of the well-known fact that, up to normalization, a likelihood ratio may be expressed as an odds ratio (Silverman and Jones, 1989). Estimating the ratio of $3b$ to $4b$ intensity functions in the Control Region thus reduces to the classification problem of estimating ψ, say by a classifier $\widehat{\psi}$. This observation has the practical advantage of circumventing the need of performing high-dimensional density estimation. Assuming that the estimator $\widehat{\psi}$ can be evaluated in the Signal Region $\mathcal{G}_s$, disjoint from its training region $\mathcal{G}_c$, we may postulate that the measure

$$A \in \mathbb{B}(\mathcal{G}_s) \longmapsto \int_A \frac{\widehat{\psi}(g)}{1 - \widehat{\psi}(g)} d\beta_3^s(g) \tag{8.14}$$

provides a reasonable approximation of β_4^s. The quality of such an approximation is driven by the ability of the classifier $\widehat{\psi}$ to generalize between regions of the phase space. To formalize this, we will assume for simplicity that $\widehat{\psi}$ is an empirical risk minimizer taking values in a class $\{\psi_\alpha : \mathcal{G} \to [0, 1] : \alpha \in \Omega\}$, for some parameter space $\Omega \subseteq \mathbb{R}^d$, $d \geq 1$. That is, we assume

$$\widehat{\psi} = \psi_{\widehat{\alpha}}, \text{ where } \widehat{\alpha} = \operatorname*{argmin}_{\alpha \in \Omega} \left\{ \frac{1}{n_c} \sum_{i=1}^{n_c} \mathcal{L}\big(\psi_\alpha(H_i^c), 0\big) + \frac{1}{m_c} \sum_{j=1}^{m_c} \mathcal{L}\big(\psi_\alpha(G_j^c), 1\big) \right\},$$

for some loss function $\mathcal{L} : [0, 1] \times [0, 1] \to \mathbb{R}$. We then make the following assumption:

Assumption 1. The conditional probability ψ satisfies the following conditions:

(i) (Correct Specification) There exists $\alpha^* \in \Omega$ such that $\psi = \psi_{\alpha^*}$.

(ii) (Generalization) We have

$$\alpha^* = \operatorname*{argmin}_{\alpha \in \Omega} \mathbb{E}\big[\mathcal{L}(\psi_\alpha(G), Z) | G \in \mathcal{G}_c\big].$$

Assumption 1 implies that a classifier trained solely in the Control Region can consistently estimate the full conditional probability $\psi(g)$, for events $g \in \mathcal{G}$ lying in both the Control and Signal Regions. Such an assumption guarantees the ability of the classifier $\widehat{\psi}$ to generalize from the Control Region, making the ansatz (8.14) justified. A natural estimator for β_4^s is then obtained by replacing β_3^s in equation (8.14) by its empirical counterpart β_{3,n_s}^s. Doing so leads to the estimator

$$\widehat{\beta}_4^s = \sum_{i=1}^{n_s} \frac{\widehat{\psi}(H_i^s)}{1 - \widehat{\psi}(H_i^s)} \delta_{H_i^s}. \tag{8.15}$$

$\widehat{\beta}_4^s$ is called the FvT estimator, and we refer to $\widehat{\psi}$ as the FvT (Four vs. Three) classifier.

The validity of Assumption 1 relies crucially upon the choice of the function class $\{\psi_\alpha\}$, or equivalently the choice of the classifier $\widehat{\psi}$. Indeed, off-the-shelf classifiers may lack the generalization ability to satisfy Assumption 1(ii). In our simulation study, we will use a classifier that is described further in Manole et al. (2022), and which has been used in the most recent $HH \to 4b$ search at the CMS Collaboration (ATLAS, 2023).

8.5 Background Modeling via Optimal Transport

The methodology described in the previous section hinged upon the ability of the classifier $\widehat{\psi}$ to accurately extrapolate from the Control Region to the Signal Region, implying that the $3b$ and $4b$ intensity functions in the latter region are constrained by their values in the former region. The validity of this assumption is difficult to verify in practice due to the blinding of the $4b$ signal region which motivates us to develop a distinct approach with a complementary modeling assumption. In this section, rather than extrapolating the discrepancy between the $3b$ and $4b$ intensity functions, we will extrapolate the discrepancy between the Control and Signal Region intensity functions, as illustrated in Figure 8.3.

We cannot use a density ratio to quantify the discrepancy between the intensity functions in the Control and Signal Regions, because these regions are disjoint. We will instead use a transport map for this purpose. In order to make use of transport maps, it will be convenient to normalize all intensity functions throughout this section. That is, we will define an estimator for β_4^s by separately estimating the probability measure $P_4^s = \beta_4^s/\beta_4^s(\mathcal{G}_s)$ and the normalization $\beta_4^s(\mathcal{G}_s)$. More generally, we denote by

$$P_j^c = \beta_j^c/\beta_j^c(\mathcal{G}_c), \quad P_j^s = \beta_j^s/\beta_j^s(\mathcal{G}_s), \quad j = 3, 4,$$

the four population-level probability measures, with corresponding empirical measures

$$P^a_{3,n_a} = \frac{1}{n_a}\sum_{i=1}^{n_a} \delta_{H_i^a}, \quad P^a_{4,m_a} = \frac{1}{m_a}\sum_{i=1}^{m_a} \delta_{G_i^a}, \quad a \in \{c, s\}.$$

Let us now formally define the optimal transport problem over $calG$.

We propose to perform background estimation under the following informal modeling assumption, which will be stated more formally in the sequel.

Assumption 2'. There exists an optimal transport map $T_0 : \mathcal{G}_c \to \mathcal{G}_s$ pushing forward P_3^c onto P_3^s such that

$$T_{0\#}P_4^c = P_4^s. \tag{8.16}$$

Assumption 2' requires the $3b$ and $4b$ distributions to be sufficiently similar for there to exist a shared map T_0 which pushes forward their restrictions to the Control Region into their counterparts in the Signal Region. If such a map T_0 were available, it would suggest the following procedure for estimating P_4^s:

(a) Fit an estimator $\widehat{T}$ of T_0 based only on the $3b$ observations;

(b) Given any estimator $\widehat{P}^c_{4,m_c}$ of P_4^c, use the pushforward $\widehat{T}_\# \widehat{P}^c_{4,m_c}$ as an estimator of P_4^s.

Let us now be more specific about the optimal transport problem arising in Assumption 2'. Unlike previous chapters, the underlying space $\mathcal{G}$ in this problem is a space of measures. We may nevertheless define the Monge problem of transporting P_3^c onto P_3^s in the usual way,

$$\underset{T:\mathcal{G}_c\to\mathcal{G}_s}{\operatorname{argmin}} \int_{\mathcal{G}_c} W(h, T(h))dP_3^c(h), \quad \text{s.t. } T_\# P_3^c = P_3^s, \tag{8.17}$$

where we take W to be a given metric on the space $\mathcal{G}$; we provide a candidate for such a metric in Section 8.5.2. When a solution T_0 to the Monge problem exists, it is said to be an optimal transport map, as in previous chapters. We postulate that the optimal transport map from P_3^c to P_3^s, when it exists, is a sensible candidate among transport maps from P_3^c to P_3^s to satisfy equation (8.16).

A shortcoming of this choice is the requirement that there exist a solution to the optimization problem (8.17). We know from Chapter 1 that the Monge problem over Euclidean space admits a unique solution for absolutely continuous distributions, when the cost function is the squared Euclidean norm. While sufficient conditions for the solvability of the Monge problem in more general spaces are given by Villani (2008, Chapter 9), we do not know whether they are satisfied by the metric space $(\mathcal{G}, W)$ under consideration. Furthermore, the Monge problem may not even be feasible between distributions which are not absolutely continuous, which precludes the possibility of estimating T_0 using the optimal transport map between the empirical measures of P_3^c and P_3^s.

Motivated by these considerations, we will prefer to formalize Assumption 2' via the Kantorovich optimal transport problem from P_3^c to P_3^s. Define the 1-Wasserstein distance over

$\mathcal{P}(\mathcal{G})$, generated by W, by

$$\mathcal{W}(P_3^c, P_3^s) = \inf_{\pi \in \Pi(P_3^c, P_3^s)} \int_{\mathcal{G}_c \times \mathcal{G}_s} W(g,h) d\pi(g,h), \tag{8.18}$$

and recall that any coupling π which achieves the infimum is referred to as an *optimal coupling* Using the Kantorovich relaxation, we now formalize Assumption 2' into the following condition, which we shall require throughout the remainder of this section.

Assumption 2. Assume there exists an optimal coupling $\pi_0 \in \Pi(P_3^c, P_3^s)$ between P_3^c and P_3^s. Given a pair of random variables $(H^c, H^s) \sim \pi_0$, let $\pi_0(\cdot|h)$ denote the conditional distribution of H^s given $H^c = h$, for any $h \in \mathcal{G}_c$. Then, the following implication holds:

$$\begin{array}{c} G^c \sim P_4^c \\ G^s | G^c \sim \pi_0(\cdot|G^c) \end{array} \implies G^s \sim P_4^s. \tag{8.19}$$

Assumption 2 requires the $3b$ and $4b$ distributions to be sufficiently similar for their restrictions to the Signal and Control Regions to be related by a common conditional distribution. It further postulates that this conditional distribution is induced by the optimal coupling π_0. Heuristically, $\pi_0(\cdot|H)$ plays the role of a multivalued optimal transport map for pushing an event H from the distribution P_3^c onto P_3^s. Assumption 2 requires this map to additionally push the distribution P_4^c onto its counterpart P_4^s in the Signal Region. In the special case where there exists an optimal transport map T_0 from P_3^c to P_3^s, we note that $\pi_0 = (Id, T_0)_\# P_3^c$ is an optimal coupling of P_3^c with P_3^s, where Id denotes the identity map. In this case, equation (8.19) is tantamount to equation (8.16).

8.5.1 Background Estimation

We next derive estimators for the background distribution P_4^s under Assumption 2. It follows from equation (8.19) and the law of total probability that

$$P_4^s(\cdot) = \int_{\mathcal{G}_c} \pi_0(\cdot|g) dP_4^c(g).$$

Since $\pi_0(\cdot|g)$ is the distribution of H^s given $H^c = g$, induced by the optimal coupling π_0, it is an identified parameter which can be estimated using only the $3b$ data. Given an estimator $\widehat{\pi}(\cdot|g)$ of this quantity, and an estimator $\widehat{P}_{4,m_c}^c$ of P_4^c, it is natural to consider the plugin estimator of the background distribution P_4^s, given by

$$\widehat{P}_4^s(\cdot) := \int_{\mathcal{G}_c} \widehat{\pi}(\cdot|g) d\widehat{P}_{4,m_c}^c(g). \tag{8.20}$$

In what follows, we begin by defining an estimator $\widehat{\pi}(\cdot|g)$ in Section 8.5.1.1, followed by two candidates for the estimator $\widehat{P}_{4,m_c}^c$, leading to two distinct background estimation methods described in Sections 8.5.1.2 and 8.5.1.3. In Section 8.5.1.4, we briefly discuss how these constructions also lead to estimators of the unnormalized intensity measure β_4^s. We then provide discussion and comparison of these methodologies in Section 8.5.1.5.

8.5.1.1 The Empirical Optimal Transport Coupling

A natural plugin estimator for the coupling π_0 is the optimal coupling $\widehat{\pi}$ between the empirical measures P^c_{3,n_c} and P^s_{3,n_s}, which has already made an appearance in 5.3 of Chapter 5. In detail, denoting by $\widehat{q} \in \mathbb{R}^{n_c \times n_s}$ the joint probability mass function of $\widehat{\pi}$, the empirical Kantorovich problem takes the following form:

$$\widehat{q} = (\widehat{q}_{ij}) \in \operatorname*{argmin}_{(q_{ij}) \in \mathbb{R}^{n_c \times n_s}} \sum_{i=1}^{n_c} \sum_{j=1}^{n_s} q_{ij} W(H^c_i, H^s_j), \ \text{ s.t. } q_{ij} \geq 0, \ \sum_{i=1}^{n_c} q_{ij} = \frac{1}{n_s}, \ \sum_{j=1}^{n_s} q_{ij} = \frac{1}{n_c}. \tag{8.21}$$

Equation (8.21) is a finite-dimensional linear program, for which exact solutions may be computed using network simplex algorithms such as the Hungarian algorithm (Kuhn, 1955). We refer to Peyré and Cuturi (2019) for a survey. We then define the estimator $\widehat{\pi}(\cdot|H^c_i)$, for $i \in [n_c]$, as the discrete distribution over $\{H^s_1, \dots, H^s_{n_s}\}$ with probability mass function

$$\widehat{q}_{j|i} = \frac{\widehat{q}_{ij}}{\sum_{k=1}^{n_s} \widehat{q}_{ik}} = n_c \cdot \widehat{q}_{ij}, \quad j = 1, \dots, n_s.$$

We are now in a position to define estimators of the background distribution P^s_4.

8.5.1.2 The OT-kNN Estimator

We first consider the general estimator in equation (8.20) when $\widehat{P}^c_{4,m_c}$ is the empirical measure P^c_{4,m_c}. This choice is perhaps most natural, but it requires us to perform out-of-sample evaluations of the estimator $\widehat{\pi}(\cdot|g)$. Indeed, recall that the latter is defined over $\{H^c_1, \dots, H^c_{n_c}\}$, whereas P^c_{4,m_c} is supported on $\{G^c_1, \dots, G^c_{m_c}\}$.

We extend the support of $\widehat{\pi}(\cdot|g)$ to all $g \in \mathcal{G}_c$ using the nearest-approach set forth in Chapter 5. Let $k \geq 1$ be an integer. For all $g \in \mathcal{G}_c$, let $I_k(g)$ denote the indices of the k-nearest neighbors of g with respect to W, among $H^c_1, \dots, H^c_{n_c}$. Specifically, we set $I(g) = \{j_1, \dots, j_k\} \subseteq [n_c]$ where

$$W(g, H^c_{j_1}) \leq \dots \leq W(g, H^c_{j_k}) \leq W(g, H^c_j), \quad \text{for all } j \in [n_c] \setminus \{j_1, \dots, j_k\}.$$

Furthermore, define the inverse distance weights

$$\omega_i(g) = \frac{1/W(g, H^c_i)}{\sum_{l \in I_k(g)} 1/W(g, H^c_l)}, \quad i \in I_k(g), \tag{8.22}$$

with the convention $\infty/\infty = 1$. We then define for all $g \in \mathcal{G}_c$,

$$\widehat{\pi}_{k\mathrm{NN}}(\cdot|g) = \sum_{i \in I_k(g)} \omega_i(g) \widehat{\pi}(\cdot|H^c_i). \tag{8.23}$$

The estimator $\widehat{\pi}_{k\mathrm{NN}}(\cdot|g)$ couples g with all of the events to which its k-nearest neighbors are coupled under $\widehat{\pi}$. The coupling values which correspond to the closest nearest neighbors are assigned higher weights $\omega_i(g)$. Furthermore, we note that when $g \in \{H_1^c, \dots, H_{n_c}^c\}$, it holds that $\widehat{\pi}_{k\mathrm{NN}}(\cdot|g) = \widehat{\pi}(\cdot|g)$. With these defintions, the generic estimator (8.20) takes the form,

$$\widehat{P}_{4,k\mathrm{NN}}^s(\cdot) := \int_{\mathcal{G}_c} \widehat{\pi}_{k\mathrm{NN}}(\cdot|g) dP_{4,m_c}^c(g) = \frac{1}{m_c} \sum_{\ell=1}^{m_c} \sum_{i \in I_k(G_\ell^c)} \omega_i(G_\ell^c) \widehat{\pi}(\cdot|H_i^c),$$

or equivalently,

$$\widehat{P}_{4,k\mathrm{NN}}^s = \frac{n_c}{m_c} \sum_{j=1}^{n_s} \left(\sum_{\ell=1}^{m_c} \sum_{i \in I_k(G_\ell^c)} \omega_i(G_\ell^c) \widehat{q}_{ij} \right) \delta_{H_j^s}.$$

We refer to $\widehat{P}_{4,k\mathrm{NN}}^s$ as the OT-kNN (Optimal Transport–k Nearest Neighbor) estimator of P_4^s.

8.5.1.3 The OT-FvT Estimator

The rate of production of $3b$ events typically exceeds that of $4b$ events by one order of magnitude (cf. Section 8.6). As a result, in the general formulation (8.20) of our optimal transport map estimators, we expect to have access to a smaller sample size m_c for estimating the distribution P_4^c, than the sample sizes n_c and n_s for estimating the optimal transport coupling π_0. Motivated by this observation, we next define an estimator $\widehat{P}_{4,m_c}^c$ which can leverage the larger $3b$ sample size n_c.

Let $p_j^c = dP_j^c / d\nu_0$ denote the density of P_j^c for $j = 3, 4$. Recall from Section 8.4 that for any event g, $\psi(g)$ denotes the probability that a random event G arose from the $4b$ distribution as opposed to the $3b$ distribution, given that $G = g$. Furthermore, $\widehat{\psi}(g)$ denotes the $[0, 1]$-valued output of the FvT classifier for discriminating $4b$ events from $3b$ events. Recall further that for any $g \in \mathcal{G}_c$, it holds that $p_4^c(g)/p_3^c(g) = (\beta_3^c(\mathcal{G}_c)/\beta_4^c(\mathcal{G}_c)) \cdot (\psi(g)/(1 - \psi(g))$, or equivalently,

$$P_4^c(A) = \frac{\beta_3^c(\mathcal{G}_c)}{\beta_4^c(\mathcal{G}_c)} \int_A \frac{\psi(h)}{1 - \psi(h)} dP_3^c(h), \quad A \in \mathbb{B}(\mathcal{G}_c).$$

We define a plugin estimator of the above quantity via

$$\widehat{P}_{4,m_c}^c(A) = \frac{n_c}{m_c} \int_A \frac{\widehat{\psi}(h)}{1 - \widehat{\psi}(h)} dP_{3,n_c}^c(h), \quad A \in \mathbb{B}(\mathcal{G}_c). \tag{8.24}$$

$\widehat{P}_{4,m_c}^c$ can be viewed as a reweighted version of the empirical measure P_{3,n_c}^c. The weights are chosen to make the $3b$ sample resemble a $4b$ sample, by using the FvT classifier to estimate the density ratio p_4^c/p_3^c. Since the $3b$ sample is one order of magnitude larger than the $4b$ sample, we heuristically expect this estimator to have smaller theoretical risk than the empirical measure P_{4,m_c}^c whenever the density ratio p_4^c/p_3^c is smooth.

A second motivation for using the estimator $\widehat{P}_{4,m_c}^c$ is the fact that it is supported on the domain of definition of the in-sample empirical optimal transport coupling $\widehat{\pi}(\cdot|g)$. We therefore do not need to extend the domain of this estimator, unlike the previous section. With these choices, the generic estimator in equation (8.20) takes the following form:

$$\widehat{P}^s_{4,\mathrm{OF}} := \int_{\mathcal{G}_c} \widehat{\pi}(\cdot|g) d\widehat{P}^c_{4,m_c}(g) = \frac{n_c}{m_c} \sum_{j=1}^{n_s} \left(\sum_{i=1}^{n_c} \frac{\widehat{\psi}(H^c_i)}{1-\widehat{\psi}(H^c_i)} \widehat{q}_{ij} \right) \delta_{H^s_j}. \tag{8.25}$$

We refer to $\widehat{P}^s_{4,\mathrm{OF}}$ as the OT-FvT (Optimal Transport–Four vs. Three) estimator of P^s_4.

8.5.1.4 Estimation of the Background Normalization

We briefly show how the OT-kNN and OT-FvT estimators can also be used to estimate the unnormalized background intensity function β^s_4. We employ the widely-used ABCD method (Alison, 2015; ATLAS, 2015; Choi and Oh, 2021; Kasieczka et al., 2021), which requires the following assumption.

Assumption 3. It holds that $\beta^s_4(\mathcal{G}_s) = \beta^s_3(\mathcal{G}_s)\beta^c_4(\mathcal{G}_c)/\beta^c_3(\mathcal{G}_c)$.

Assumption 3 implies that the ratio of the number of $4b$ to $3b$ events in the Control Region should be the same as that in the Signal Region. Under this assumption, a natural estimator for $\beta^s_4(\mathcal{G}_s)$ is simply given by $m_c n_s/n_c$. Therefore, under Assumptions 2–3, the probability measures $\widehat{P}^s_{4,k\mathrm{NN}}$ and $\widehat{P}^s_{4,\mathrm{OF}}$ can be used to define the following two estimators of the unnormalized background intensity measure β^s_4,

$$\widehat{\beta}^s_{4,k\mathrm{NN}} = \frac{m_c n_s}{n_c} \widehat{P}^s_{4,k\mathrm{NN}}, \quad \widehat{\beta}^s_{4,\mathrm{OF}} = \frac{m_c n_s}{n_c} \widehat{P}^s_{4,\mathrm{OF}}. \tag{8.26}$$

We respectively refer to the above measures as the OT-kNN and OT-FvT estimators of β^s_4, or simply as the OT-kNN and OT-FvT methods.

8.5.1.5 Remarks

We summarize the three background estimation methods, FvT, OT-kNN and OT-FvT, in Table 8.1, and make the following remarks:

- Assumption 2 is the primary modeling assumption required by OT-kNN and OT-FvT. We view this condition as being complementary to Assumption 1(ii), required by the FvT method. Indeed, it involves an extrapolation (of an optimal coupling) from the $3b$ to $4b$ distribution, rather than an extrapolation (of a density ratio) from the Control Region to the Signal Region.
- The OT-FvT estimator (8.25) can alternatively be interpreted through the lens of domain adaptation for the FvT classifier. To make this connection clear, suppose for simplicity that $n_c = m_c$. In this case, it can be shown that $\widehat{\pi}$ is in fact induced by an optimal transport map, in the sense that there exists a permutation $\widehat{\tau} : [n_c] \to [n_c]$ such that

$$\widehat{q}_{ij} = I(i = \widehat{\tau}(j))/n_c, \quad i, j = 1, \dots, n_c.$$

Table 8.1: Summary of the three background estimation methods: FvT, OT-kNN, and OT-FvT. The final estimator for each method takes the form $\widehat{\beta}_4^s \propto \sum_{j=1}^{n_s} v_j \delta_{H_j^s}$, for the values of v_j listed in the table.

Estimator (of the form $\propto \sum_{j=1}^{n_s} v_j \delta_{H_j^s}$)	FvT	OT-FvT	OT-kNN
v_j	$\dfrac{\widehat{\psi}(H_j^s)}{1-\widehat{\psi}(H_j^s)}$	$\displaystyle\sum_{i=1}^{n_c} \frac{\widehat{\psi}(H_i^c)}{1-\widehat{\psi}(H_i^c)} \widehat{q}_{ij}$	$\displaystyle\sum_{\ell=1}^{m_c} \sum_{i \in I_k(G_\ell^c)} \omega_i(G_\ell^c) \widehat{q}_{ij}$

The FvT and OT-FvT estimators then take the following form:

$$\widehat{\beta}_{4,\mathrm{FvT}}^s \propto \sum_{j=1}^{n_s} \frac{\widehat{\psi}(H_j^s)}{1-\widehat{\psi}(H_j^s)} \delta_{H_j^s}, \qquad \widehat{\beta}_{4,\mathrm{OF}}^s \propto \sum_{j=1}^{n_s} \frac{\widehat{\psi}(H_{\widehat{\tau}(j)}^c)}{1-\widehat{\psi}(H_{\widehat{\tau}(j)}^c)} \delta_{H_j^s}.$$

While the FvT method evaluates the density ratio estimator $\widehat{\psi}/(1-\widehat{\psi})$ at events H_j^s in the Signal Region, the OT-FvT method evaluates it at the events $H_{\widehat{\tau}(j)}^c$ in the Control Region, to which the events H_j^s are mapped under the empirical optimal coupling $\widehat{\pi}$. The OT-FvT method thus circumvents the evaluation of $\widehat{\psi}$ outside the region where it was trained. Optimal transport has similarly been used in past literature as a tool for domain adaptation between train and test data in classification problems (cf. Section 8.1.2).

- In defining the estimator OT-kNN, we proposed to extend the domain of definition of the empirical optimal transport coupling $\widehat{\pi}(\cdot|g)$ to the entire space $\mathcal{G}_c$ via nearest neighbor extrapolation; cf. equation (8.23). We saw in Chapter 5 that, for the quadratic optimal transport problem over Euclidean space, such a procedure has statistically minimax optimal risk for estimating the underlying optimal transport map T_0, assuming that it exists and is Lipschitz continuous. Nevertheless, the risk of this estimator suffers severely from the curse of dimensionality, and does not generally improve when T_0 enjoys higher regularity. We instead showed in Chapter 5 that plugin estimators of T_0 based on density estimates of P_3^c and P_3^s may achieve improved convergence rates in such settings. In our context, it is challenging to perform density estimation over the space of measures $\mathcal{G}$—and particularly over the non-convex set $\mathcal{G}_c$—thus we did not follow this approach. Our aim was instead to alleviate the curse of dimensionality inherent to the OT-kNN method by introducing the OT-FvT method. Indeed, we view the task of estimating P_4^c as a larger statistical bottleneck than that of estimating π_0, and the estimator $\widehat{P}_{4,m_c}^c$ (used by the OT-FvT method) may potentially achieve smaller risk than the empirical measure P_{4,m_c}^c (used by the OT-kNN method).
- Our theory in Chapter 5 additionally shows that the value $k = 1$ suffices for the estimator $\widehat{\pi}_{k\mathrm{NN}}$ to enjoy optimal theoretical risk, at least for the Euclidean optimal transport problem with quadratic cost. In our work, we nevertheless allow for k to be greater than 1 in order to leverage the larger size of the $3b$ sample. For example, when $k = 1$, the estimator $\widehat{\beta}_{4,k\mathrm{NN}}^s$ is supported on at most m_c events, whereas it can be supported on as

many as $n_s \gg m_c$ events if k is chosen sufficiently large. In practice, we recommend choosing k to be as small as possible while ensuring that $\widehat{\beta}^s_{4,k\mathrm{NN}}$ has support size on the same order as n_s—this typically amounts to choosing k to be on the order of n_s/m_c. In our simulation study (cf. Section 8.6), we therefore choose the value $k = 10$, but also illustrate the performance of the OT-kNN method for other values of k.

- We have chosen to separately estimate the probability measure P^s_4 and the normalization $\beta^s_4(\mathcal{G}_s)$, because the classical optimal transport problem is only well-defined between measures with the same total mass. A possible alternative is to consider the *partial* (Figalli, 2010) or *unbalanced* (Liero, Mielke, and Savaré, 2018) optimal transport problems between the unnormalized intensity measures β^c_3 and β^s_3. These variants of optimal transport are well-defined between measures that have possibly different mass, but have the downside of introducing tuning parameters. As we explain in Section 8.6, the normalizations $\beta^c_3(\mathcal{G}_c)$ and $\beta^s_3(\mathcal{G}_s)$ are of the same order of magnitude, and can in fact be made to coincide by tuning the definition of the Control and Signal regions, thus we have simply focused our attention on the classical (balanced) optimal transport problem in this work. Nevertheless, in the following subsection, we will employ a variant of the partial optimal transport problem to define the metric W.

8.5.2 A Metric between Collider Events

We now describe a candidate for the metric W on $\mathcal{G}$. Recall that the Kantorovich problem in (8.18) gave rise to the Wasserstein distance $\mathcal{W}$ between probability distributions over $\mathcal{G}$. By a recursion of ideas, we will also define W to be a Wasserstein-type metric, arising from the optimal transport problem between constituent jets of events. This approach was introduced by Komiske, Metodiev, and Thaler (2019). They propose to metrize $\mathcal{G}$ using a variant of the Wasserstein distance which is well-defined between measures with non-equal mass (Peleg, Werman, and Rom, 1989; Pele and Werman, 2008). Given any two collider events $g = \sum_{j=1}^4 p_{T_j}\delta_{(\eta_j,\phi_j,m_j)} \in \mathcal{G}$, $h = \sum_{j=1}^4 p'_{T_j}\delta_{(\eta'_j,\phi'_j,m'_j)} \in \mathcal{G}$, the metric is defined by

$$\begin{aligned}\widetilde{W}(g,h) = \min_{(f_{ij})\in\mathbb{R}^{4\times4}} \frac{1}{R}\sum_{i=1}^4\sum_{j=1}^4 f_{ij}\sqrt{(\eta_i-\eta'_j)^2+(\phi_i-\phi'_j)^2} + \left|\sum_{i=1}^4 (p_{T_i}-p'_{T_i})\right| \\ \text{s.t.}\quad f_{ij}\geq 0,\quad \sum_j f_{ij}\leq p_{T_i},\quad \sum_i f_{ij}\leq p'_{T_j},\quad \sum_{i,j} f_{ij} = \min(\sum_i p_{T_i}, \sum_j p'_{T_j}),\end{aligned} \tag{8.27}$$

for a tuning parameter $R > 0$. We make several remarks about this definition.

- In the context of particle physics, the coupling f_{ij} is naturally interpreted as a flow of energy (measured in terms of the transverse momentum p_T) from jet i of g to jet j of h, as depicted in Figure 8.4. $\widetilde{W}(g,h)$ thus measures the smallest possible transport of energy required to rearrange the jets of the event g into those of h.
- We have followed Komiske, Metodiev, and Thaler (2019) by omitting the mass variables m_j and m'_j from the definition of $\widetilde{W}$. This choice is further discussed in the context of

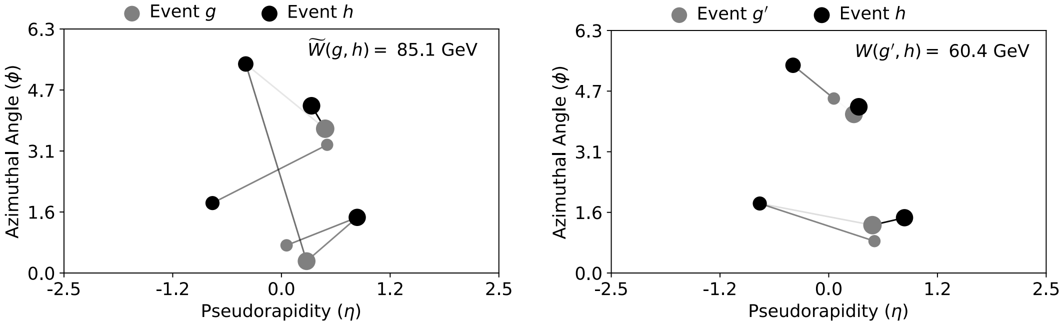

Figure 8.4: Left: (η, ϕ)-plot of two events $g, h \in \mathcal{G}$. Each point represents a constituent jet, with size proportional to its p_T value. A line connecting the i-th jet of event g to the j-th jet of event h indicates a nonzero value of the optimal coupling f_{ij}, with line darkness increasing as a function of the magnitude of f_{ij}. Right: (η, ϕ)-plot of events $g', h \in \mathcal{G}$, where $g' \simeq g$ is an approximate minimizer in Eq. (8.28). One has $W(g,h) = W(g',h) < \widetilde{W}(g,h)$.

our simulation study in Section 8.6.

- The tuning parameter R trades-off the influence of the angular variables ϕ_i, η_i, and that of the energy variables p_{T_i}. Our choice of R is further discussed in Section 8.6.

The metric $\widetilde{W}$ does not, however, take into account the equivalence relation $\simeq$ over $\mathcal{G}$ defined in equation (8.2). For example, $\widetilde{W}(g,h)$ could be nonzero even when g and h are deemed equivalent for our purposes. We therefore define our final metric W by

$$W(g,h) = \inf\left\{\widetilde{W}(g',h) : g' \simeq g,\ g' \in \mathcal{G}\right\}, \quad g,h \in \mathcal{G}. \tag{8.28}$$

Strictly speaking, W now becomes a metric over the set of equivalence classes of events induced by $\simeq$. We refer to Figure 8.4 for an illustration.

8.6 Simulation Study

8.6.1 Simulation Description

In this section, we compare the performance of the three background modeling methods OT-FvT, OT-kNN and FvT, on realistic simulated collider data, generated using the MadGraph particle physics software (Alwall et al., 2011). Code for reproducing this simulation study is publicly available[4].

Since b-tagging is imperfect, in practice, we expect the $3b$ and $4b$ samples to be composed of a mixture of different multijet scattering processes which do necessarily arise from b-quarks.

[4] https://github.com/tmanole/HH4bsim

We perform a study in MadGraph to estimate the relative scale of such processes. Assuming a b-jet tagging efficiency of 75%, a charm jet tagging efficiency of 15% and a light jet tagging efficiency of 1%, we find that the $4b$ (resp. $3b$) sample consists of 90% (10%) events in a final state with four b quarks, 7% (9%) events in a final state with two b quarks and two charm quarks, and 4% (80%) in a final state with two b quarks and two light quarks. In particular, we stress that a fraction of the $3b$ sample consists of mislabelled $4b$ events, which could be signal events. This signal contamination is expected to be sufficiently small to be considered negligible for purposes of a signal discovery analysis, as in this chapter.

We generate four-quark events in MadGraph according to the percentages listed above. The calorimeters in the CMS detector are not perfect, and the measured jet energies have a finite resolution. The distribution of the observed smeared energy is well-approximated by the normal distribution $N(E, \sigma^2(E))$, where E denotes the true energy of a jet, and $\sigma(E)$ satisfies

$$\left(\frac{\sigma(E)}{E}\right)^2 = \left(\frac{S}{\sqrt{E}}\right)^2 + \left(\frac{N}{E}\right)^2 + C^2,$$

for some constants $S, N, C \geq 0$. We apply this smearing to the quark four-vectors, setting $S = 0.98, N = 0, C = 0.054$. For simplicity we set the quark masses to zero and omit them from the the metric W. When applying these methods to real data, it may, however, be useful to incorporate the jet masses into the definition of W. We also apply jet-level scale factors to account for the p_T dependence of CMS b-tagging for light, charm and bottom quark jets:

$$\text{Scale Factor} = \begin{cases} (2.5\, p_T\, e^{-7\, p_T} + 0.6)/0.75 & b\text{-quark} \\ (p_T\, e^{-10\, p_T} + 0.2)/0.15 & c\text{-quark} \\ (0.03\, p_T + 0.01)/0.01 & u, d, s\text{-quark or gluon,} \end{cases}$$

where p_T is measured in TeV. Events are weighted by the product of the scale factors for the b-tagged jets.

Following this pre-processing of the data, the pairing algorithm described in Section 8.3.2 is applied to all events, and those falling within the Control and Signal Regions are kept. We define these regions according to equations (8.9–8.10), with the parameters $\sigma_c = 1.03$, $\kappa_s = 1.6$ and $r_c = 30\,\text{GeV}$. The final sample consists of $n_s = 201,568$ events in the $3b$ Signal Region, $n_c = 159,427$ events in the $3b$ Control Region, $m_s = 28,980$ events in the 4b Signal Region, and $m_c = 22,053$ events in the 4b Control Region. The order of magnitude of these sample sizes, as well as the proportion of $3b$ to $4b$ events, is similar to those used in recent di-Higgs analyses at the LHC (ATLAS, 2019). We also simulate a separate 4b sample of size approximately $10(n_s + m_s)$, which we choose not to contain any signal events, and whose distribution we treat as the ground truth for the purpose of validating our background models.

We additionally generate a Monte Carlo sample from the SM di-Higgs signal distribution, with which the signal intensity rates $(S_j)_{j=1}^J$, used to form the likelihood function (8.7), can be specified. For the purpose of validating our background models, we train a $[0, 1]$-valued classifier $\widehat{\xi}$ (abbr. Signal vs. Background, or SvB, classifier) to discriminate the $4b$ data from the Monte Carlo di-Higgs sample. Given that our simulated 4b sample contains no signal events, $\widehat{\xi}$

forms a reasonable proxy for the theoretical binning function ξ in equation (8.8). The SvB classifier has the same architecture as that of the FvT classifier described in in Manole et al. (2022). In the sequel, we refer to $\widehat{\xi}(g)$ as the SvB value corresponding to an event g.

Finally, we discuss our choice of the parameter R arising in the definition of the metric W. In order for the two terms in the definition of $\widetilde{W}$ to be of comparable order, we make the requirement that R lie within the range of the first summand in equation (8.27). We identify this range as follows. Since b-tagging is only performed for values of η lying in the interval $[-2.5, 2.5]$, we impose $R \leq \sqrt{\pi^2 + 5^2} \approx 5.9$. Furthermore, jet clustering algorithms used by CMS merge particles whose (η, ϕ)-Euclidean distance is within 0.4 (Cacciari, Salam, and Soyez, 2008; CMS, 2017), thus we impose $R \geq 0.4$. Now, since we expect that the largest discrepancies between the Control and Signal Region distributions arise in the kinematic variables (η, ϕ), we choose the smallest possible value $R = 0.4$ when fitting the empirical optimal transport coupling $\widehat{\pi}$. On the other hand, for the nearest-neighbor lookup of the OT-kNN method, we set $R = 2.75$, which is the midpoint of the interval $[0.4, 5.9]$. We make no attempt to tune these values of R, and we leave open the question of choosing them in a data-driven fashion. We compute the metric W in part using the EnergyFlow Python library (Komiske, Metodiev, and Thaler, 2022), and we compute optimal couplings between distributions of collider events using the Python Optimal Transport library (Flamary et al., 2021)—see Appendix 8.A for further details.

8.6.2 Simulation Results

The fitted intensity measures $\widehat{\beta}_4^s$ produced by the three background methods (FvT, OT-kNN and OT-FvT) are binned and plotted in Figure 8.5. Logarithmic scales are used to better visualize signal-rich regions. It can be seen that the three methods yield qualitatively similar estimates of the SvB intensity function. We recall that the SvB variable is of primary interest to model, as it is used as the final discriminant when testing the signal hypothesis (8.3). The m_{HH} variable has also been used as the final discriminant in recent di-Higgs studies (Bryant, 2018). Given an event $g \in \mathcal{G}$ with dijet pairing $\{g^1, g^2\}$, its m_{HH} value is defined as follows[5], using the notation of Section 8.3.2,

$$m_{\mathrm{HH}}(g) = m\left(\frac{m_{\mathrm{H}}}{m(g^1)}g^1 + \frac{m_{\mathrm{H}}}{m(g^2)}g^2\right). \tag{8.29}$$

Once again, we observe that this variable is well-modelled by all three methods.

In order to provide a quantitative comparison of these methods, we develop a heuristic two-sample test for testing equality between the distribution of the fitted background models and of the true upsampled $4b$ data. Specifically, we form a proxy for a two-sample test by training classifiers to discriminate each of the background estimates from the upsampled $4b$ data (similar approaches have previously used in the high energy physics literature by Krause and Shih (2023a, 2023b). For each classifier, we record the area under the receiver operating characteristic curve (AUC; Hanley and McNeil (1982)), and any deviation of this quantity from .5

[5] Equation (8.29) can be interpreted as the four-body invariant mass after the dijet four-vectors have been corrected to have the Higgs boson mass.

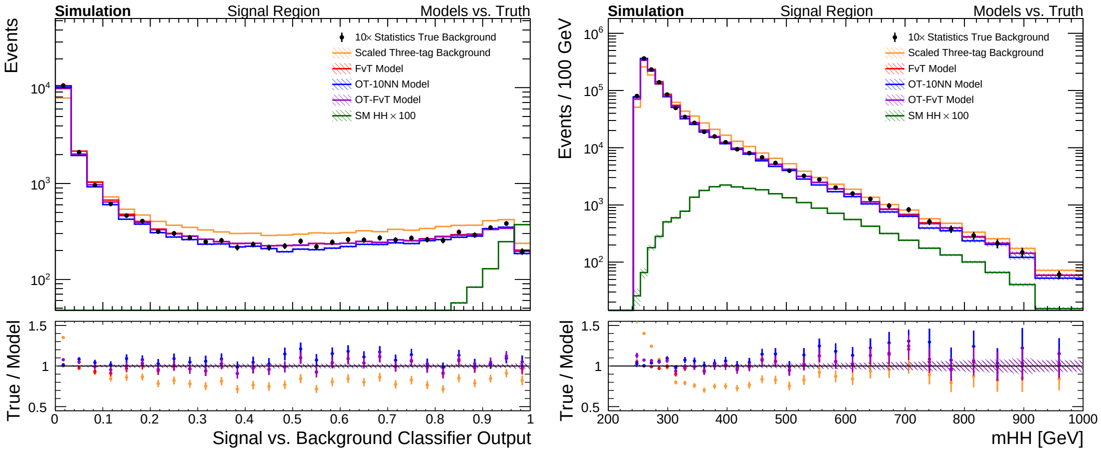

Figure 8.5: Histograms of the the SvB classifier output variable (left) and the m_{HH} variable (right) for the three background models as well as the upsampled 4b data (treated as the ground truth), the 3b data (normalized by the factor $n_s m_c/n_c$), and the di-Higgs signal sample (SM HH). Error bars in the k-th bin of any histogram denote $\pm\sqrt{N_k}$, where N_k is the number of events per bin. Error bars in the ratio plot denote $\pm\sqrt{N_k}/N_{0k}$, where N_{0k} is the number of observed $4b$ events per bin. The dashed lines in the ratio plot denote $\pm\sqrt{1/N_{0k}}$.

is an indication of mismodeling. We again choose our classifiers to be residual neural networks with the architecture described in Manole et al. (2022). Although this choice is inherently favourable to the FvT method, and to some extent the OT-FvT method, we use it because it coincides with the SvB classifier architecture, and will thus be most powerful at detecting mismodeling in the features which are relevant for the final signal analysis. Small sentence to address point #6 from the referee — let me know if we should say more. Another caveat with the use of the AUC as a performance metric is the fact that it may not be sensitive to local deviations in the signal-rich area of the phase space, since it has low overall yield. The fitted AUC value for each method is reported in Figure 8.6. Though all AUCs are significantly greater than .5, they are substantially lower for our background models than for the benchmark method consisting of the uncorrected 3b sample. The FvT method has the lowest AUC, followed closely by the OT-FvT method and OT-kNN method. While the OT-1NN method has comparable AUC point estimate as the OT-FvT method, we emphasize that its variability interval is wider, which could have been anticipated from the discussion in Section 8.5.1.5, where we emphasized that the support size of $\beta^s_{4,1\mathrm{NN}}$ can be an order of magnitude smaller than n_s. In contrast, the OT-10NN and OT-20NN estimators have narrower variability intervals than OT-1NN, but have markedly larger AUC point estimates than the remaining methods. The performance of the OT-kNN method for varying values of k is also illustrated in Figure 8.8 along as a function of the SvB and m_{HH} variables.

We next provide a qualitative comparison of the fitted weights produced by the three

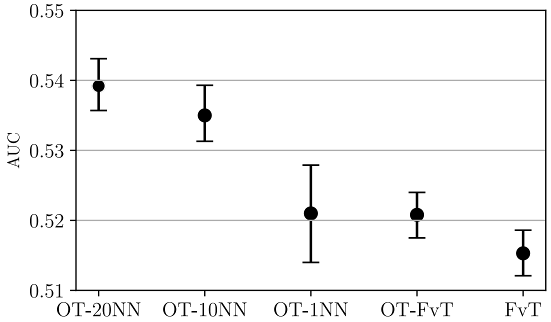

Figure 8.6: Fitted AUC values obtained by discriminating each background model from the upsampled 4b data using the FvT classifier, together with 95% percentile bootstrap variability intervals, obtained by bootstrapping the predicted classifier probabilities. 1,000 bootstrap replications are used. Note that this bootstrap procedure does not take into account the variability of the background estimators themselves. For the $3b$-tagged data, we obtain the AUC 0.5843, with variability interval [0.5812,0.5874].

background modeling methods. Recall that these methods all take the form

$$\widehat{\beta}_4^s = \sum_{i=1}^{n_s} v_i \delta_{H_i^s},$$

for some nonnegative weights v_i, which are summarized up to normalization in Table 8.1. In Figure 8.8, we plot the weights of the two optimal transport methods against those of the FvT method. We observe that the FvT and OT-FvT methods produce weights which are concentrated and symmetric around the identity. This implies that the odds ratio of the FvT classifier at a point H_j^s in the Signal Region behaves similarly to the odds ratio at any point H_i^c in the Control Region to which H_j^s is optimally coupled. This suggests that the transfer learning of the FvT classifier from the Control Region to the Signal Region is, to some extent, well-modelled by the optimal transport coupling $\widehat{\pi}$. This observation heuristically suggests that Assumptions 1–2 both hold in this simulation. In contrast to the method OT-FvT, we observe that the method OT-10NN produces markedly different weights than the FvT method, which can partly be anticipated from the discrete nature of the nearest neighbor extrapolation. We conjecture that the nearest-neighbor estimator of the optimal transport coupling has poorer theoretical risk than its counterpart in the OT-FvT method.

8.7 Conclusion and Discussion

Our aim has been to study the problem of data-driven background estimation, motivated by the ongoing search for double Higgs boson production in the $4b$ final state. After recalling a widely-used approach to this problem based on transfer learning of a multivariate classifier, our first contribution was to develop the FvT classifier architecture which is tailored to collider

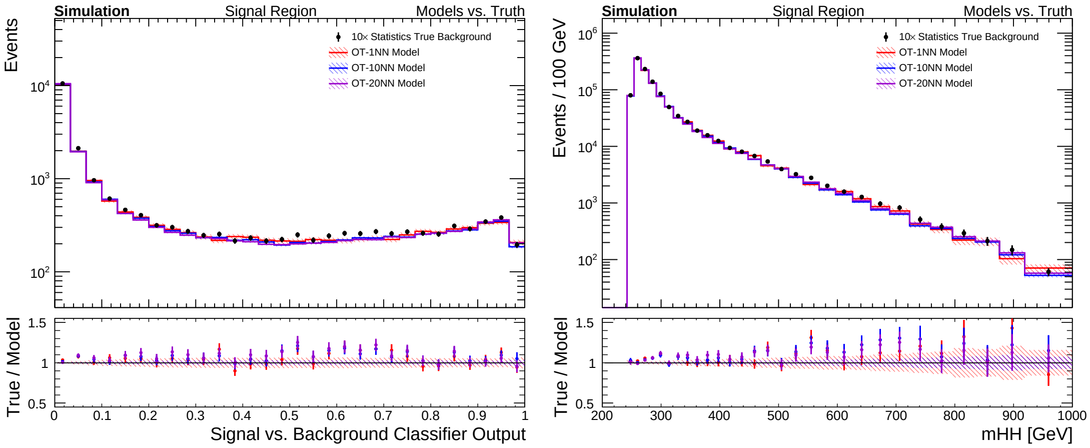

Figure 8.7: Histograms of the SvB classifier output variable (left) and the m_{HH} variable (right) for the OT-kNN estimator, with $k \in \{1, 10, 20\}$.

data, and which can serve as a powerful tool for implementing this methodology. Our primary contribution was then to propose a distinct background estimation method based on the optimal transport problem. A recurring theme throughout our work has been the complementarity of the modeling assumptions made by these two distinct approaches, which allows them to be used as cross-checks for one another in practice. We substantiate this point with a realistic simulation study, in which these methodologies appear to give consistent results despite their inherently distinct derivations.

Quantifying the uncertainty of our background estimates is a challenging problem left open by our work, which is nonetheless crucial for applying our methods in practice. In the experimental particle physics community, it is commonplace to measure both *statistical* uncertainties—those arising from fluctuations of the data generating process—and *systematic* uncertainties—those arising from potential mismodeling (Heinrich and Lyons, 2007). Both of these forms of uncertainty are challenging to quantify in our context. For instance, a prerequisite for quantifying the statistical uncertainty of the methods OT-kNN or OT-FvT is to obtain uniform confidence bands for optimal transport maps. Although we have taken a first step toward this question in Chapter 6, our theory is limited to pointwise confidence bands over the torus, and completely new ideas would be needed to obtain practical uniform bands over the space of measures $\mathcal{G}$. The question of quantifying systematic uncertainties is more open-ended, and typically involves heuristics for assessing the extent of potential mismodeling by the background estimation methods. Due to the complementarity of assumptions placed by our methods, any lack of closure between them could potentially play a role in quantifying their systematic uncertainties. While further investigation is required to make such a proposal formal, it is our hope that the optimal transport methodology presented in our work can help contribute to the challenging question of systematic uncertainty quantification in the search for di-Higgs boson production, or in other searches at the Large Hadron Collider.

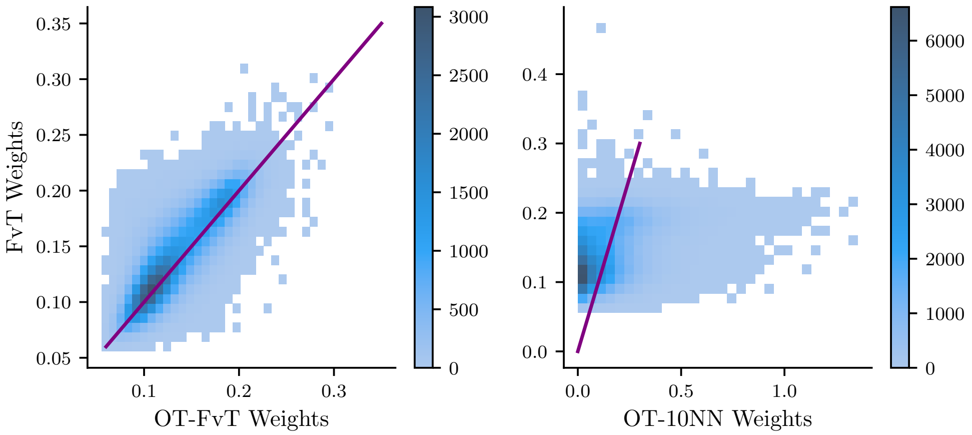

Figure 8.8: Bivariate histogram of the $3b$ data $H_1^s, \dots, H_{n_s}^s$ in the Signal Region, plotted in terms of the weights of the OT-FvT method against those of the FvT method (left), and of the OT-10NN method against those of the FvT method (right). The purple lines denote the identity function.

8.A Computation of Optimal Transport Couplings

In this section, we describe our numerical approximation of the empirical optimal transport coupling $\widehat{q}$ in equation (8.21). Equation (8.21) is a linear program which can be computed exactly using simplex algorithms (Peyré and Cuturi, 2019). Such approaches have memory complexity which grows quadratically in $n_c \wedge n_s$, since they require the cost matrix

$$C = (W(H_i^c, H_j^s) : 1 \le i \le n_c, 1 \le j \le n_s)$$

to be stored in memory. As described in Section 8.6, the sample sizes n_c and n_s are at least of the order 10^5 in our problem, in which case the storage of the matrix C becomes intractable. For low-dimensional problems, the storage of C can be avoided by using the so-called back-and-forth algorithm of Jacobs and Léger (2020), which has linear memory complexity. For higher dimensional problems, a common approach is to divide the datasets into several (say, B) batches, and to compute B separate optimal transport couplings. Such batches can either be obtained through subsampling or deterministic schemes—see Sommerfeld et al. (2019), Fatras et al. (2021), Nguyen et al. (2022), and references therein.

We follow a similar approach in our work. We partition the two samples

$$\mathcal{D}^c = \{H_1^c, \dots, H_{n_c}^c\}, \quad \mathcal{D}^s = \{H_1^s, \dots, H_{n_s}^s\}$$

into $B \ge 1$ disjoint batches $\mathcal{D}_1^c, \dots, \mathcal{D}_B^c$ and $\mathcal{D}_1^s, \dots, \mathcal{D}_B^s$, satisfying $\mathcal{D}^c = \bigcup_k \mathcal{D}_k^c$ and $\mathcal{D}^s = \bigcup_k \mathcal{D}_k^s$. Assume for simplicity that for some $n_B^c, n_B^s \ge 1$, $|\mathcal{D}_k^c| = n_c^B$ and $|\mathcal{D}_k^s| = n_s^B$ for all $k = 1, \dots, B$. We then compute the optimal transport couplings

$$\widehat{q}^k = \operatorname*{argmin}_{(q_{ij}) \subseteq \mathbb{R}^{n_c^B \times n_s^B}} \sum_{H^c \in \mathcal{D}_k^c} \sum_{H^s \in \mathcal{D}_k^s} q_{ij} W(H^c, H^s)$$

where the minimum is taken over all $q_{ij} \geq 0$ satisfying

$$\sum_{i=1}^{n_c^B} q_{ij} = \frac{1}{n_s^B}, \quad \sum_{j=1}^{n_s^B} q_{ij} = \frac{1}{n_c^B},$$

for all $k = 1, \dots, B$. We then approximate the empirical optimal transport coupling $\widehat{q}$ in equation (8.21) by the matrix

$$\widetilde{q} = \frac{1}{B} \begin{pmatrix} \widehat{q}^1 & & & \\ & \widehat{q}^2 & & \\ & & \ddots & \\ & & & \widehat{q}^B \end{pmatrix}.$$

The memory complexity of this algorithm is $O(Bn_c^B n_s^B)$ as opposed to the complexity $O(n_c n_s)$ incurred by any method which requires the storage of the matrix C.

In our simulations, we computed the couplings $\widehat{q}_k$ using the network simplex solver described by Bonneel et al. (2011), as implemented in the Python Optimal Transport package (Flamary et al., 2021). Our simulations were run on a standard Linux machine with 12 cores and 32GB of RAM. We chose $B = 16$, which is approximately the smallest value of B for which the computation of the couplings $\widehat{q}^k$ did not exceed our machine's memory limit.

We chose the batches according to the following procedure. For any event $G = \sum_{j=1}^4 p_{T_j} \delta_{(\eta_j, \phi_j, m_j)}$, let $s_T(G) = \sum_{j=1}^4 p_{T_j}$ denote the scalar sum of the transverse momenta of G. Let $H_{(i)}^c$ denote the event among $H_1^c, \dots, H_{n_c}^c$ with the i-th smallest s_T value, for all $i = 1, \dots, n_c$. That is,

$$s_T(H_{(1)}^c) \leq s_T(H_{(2)}^c) \leq \cdots \leq s_T(H_{(n_c)}^c).$$

We likewise define $H_{(j)}^s$ for $j = 1, \dots, n_s$ such that

$$s_T(H_{(1)}^s) \leq s_T(H_{(2)}^s) \leq \cdots \leq s_T(H_{(n_s)}^s).$$

We then set, for $k = 1, \dots, B$,

$$\mathcal{D}_k^c = \{H_{(B(r-1)+k)}^c : 1 \leq r \leq n_c^B\}, \quad \mathcal{D}_k^s = \{H_{(B(r-1)+k)}^s : 1 \leq r \leq n_s^B\}.$$

This choice ensures that each batch contains events with a comparable range of s_T values. We impose this property because the penalty term in the definition of $\widetilde{W}$ (Eq. (8.27)) is sensitive to large deviations of s_T values. We leave open for future work whether different batching methods, such as subsampling, would yield improved performance.

Appendix A

Function Spaces

In this Appendix, we collect definitions and properties of Hölder spaces, Besov spaces, and Sobolev Spaces, which are used throughout this thesis.

A.1 Hölder Spaces

Given a closed set $\Omega \subseteq \mathbb{R}^d$, let $\mathcal{C}_u(\Omega)$ denote the set of uniformly continuous real-valued functions on Ω. For any function $f : \Omega \to \mathbb{R}$ which is differentiable up to order $k \geq 1$ in the interior of Ω, and any multi-index $\gamma \in \mathbb{N}^d$, we write $|\gamma| = \sum_{i=1}^d \gamma_i$, and for all $|\gamma| \leq k$,

$$D^\gamma f = \frac{\partial^{|\gamma|} f}{\partial x_1^{\gamma_1} \dots \partial x_d^{\gamma_d}}.$$

Given an integer $k \geq 0$ and $\beta \in (0,1)$, the Hölder space $\mathcal{C}^{k,\beta}(\Omega)$ is defined as the set of functions $f \in \mathcal{C}_u(\Omega)$ which are differentiable to order k in the interior of Ω, with derivatives extending continuously up to the boundary of Ω, and such that the Hölder norm

$$\|f\|_{\mathcal{C}^{k,\beta}(\Omega)} = \sum_{j=0}^{k} \sup_{|\gamma|=j} \|D^\gamma f\|_\infty + \sum_{|\gamma|=k} \sup_{\substack{x,y\in\Omega^\circ \\ x\neq y}} \frac{|D^\gamma f(x) - D^\gamma f(y)|}{\|x-y\|^\beta}$$

is finite. We typically use the abbreviation

$$\mathcal{C}^\alpha(\Omega) := \mathcal{C}^{\lfloor \alpha \rfloor, \alpha - \lfloor \alpha \rfloor}(\Omega),$$

for any $\alpha > 0$, and in Chapter 3 we will also use the abbreviation

$$\mathscr{C}^\alpha(\Omega) := \mathcal{C}^{\underline{\alpha}, \alpha - \underline{\alpha}}(\Omega),$$

where $\underline{\alpha}$ denotes the largest integer strictly less than α (for instance, $\underline{\alpha} = 0$ when $\alpha = 1$). Furthermore, for any $\alpha \geq 0, \mathcal{C}^\alpha(\mathbb{T}^d)$ (resp. $\mathcal{C}_u(\mathbb{T}^d)$) is defined as the set of $\mathbb{Z}^d$-periodic functions $f : \mathbb{R}^d \to \mathbb{R}$ such that $f \in \mathcal{C}^\alpha(\mathbb{R}^d)$ (resp. $f \in \mathcal{C}_u(\mathbb{R}^d)$).

Let us now state a few elementary facts about Hölder spaces. The following simple interpolation inequality can be found, for instance, in equations (6.8)–(6.9) of Gilbarg and Trudinger (2001).

Lemma 94 (Interpolation Inequality for Hölder Norms). *For all $\epsilon > 0$, there exists $C(\epsilon, d) > 0$ such that for all $f \in \mathcal{C}^{2+\beta}(\mathbb{T}^d)$,*

$$\|f\|_{\mathcal{C}^{1+\beta}(\mathbb{T}^d)} \leq C(\epsilon)\|f\|_{L^\infty(\mathbb{T}^d)} + \epsilon\|f\|_{\mathcal{C}^{2+\beta}(\mathbb{T}^d)}.$$

The following bound can be deduced from the Leibniz rule together with equation (4.7) of Gilbarg and Trudinger (2001).

Lemma 95 (Products of Hölder Functions). *For any $\alpha \geq 0$, there exists a constant $C = C(\alpha) > 0$ such that for all $f, g \in \mathcal{C}^\alpha(\mathbb{T}^d)$,*

$$\|fg\|_{\mathcal{C}^\alpha(\mathbb{T}^d)} \leq C_\alpha\|f\|_{\mathcal{C}^\alpha(\mathbb{T}^d)}\|g\|_{\mathcal{C}^\alpha(\mathbb{T}^d)}.$$

The following is straightforward.

Lemma 96 (Compositions of Hölder Functions). *For any $\alpha, \beta \in (0, 1]$, there exists a constant $C = C(\alpha, \beta) > 0$ such that for all $f \in \mathcal{C}^{1+\alpha}(\mathbb{T}^d)$ and $g \in \mathcal{C}^\beta(\mathbb{T}^d)$,*

$$\|g \circ \nabla f\|_{\mathcal{C}^{\alpha\beta}(\mathbb{T}^d)} \leq C\|f\|^\beta_{\mathcal{C}^{1+\alpha}(\mathbb{T}^d;\mathbb{T}^d)}\|g\|_{\mathcal{C}^\beta(\mathbb{T}^d)}.$$

A.2 Sobolev Spaces

A.2.1 Sobolev Spaces on the Torus

Fix the collection of test functions $\mathcal{D}(\mathbb{T}^d) = \mathcal{C}^\infty(\mathbb{T}^d)$, endowed with the standard test function topology, and let $\mathcal{D}_0(\mathbb{T}^d) = \mathcal{C}_0^\infty(\mathbb{T}^d)$. The set of periodic distributions $\mathcal{D}'(\mathbb{T}^d)$ is defined as the set of continuous linear functionals on $\mathcal{C}^\infty(\mathbb{T}^d)$, and we denote by $\langle \cdot, \cdot \rangle$ the induced duality pairing. $\mathcal{D}'_0(\mathbb{T}^d)$ is similarly defined as the dual of $\mathcal{D}_0(\mathbb{T}^d)$. Furthermore, define the discrete Schwartz space $S(\mathbb{Z}^d)$ as the set of maps $\phi : \mathbb{Z}^d \to \mathbb{R}$ such that for any $k > 0$ there exists $C_k > 0$ such that

$$|\phi(\xi)| \leq C_k\|\xi\|^{-k}, \quad \text{for all } \xi \in \mathbb{Z}^d.$$

The set of tempered distributions on $\mathbb{Z}^d$ is denoted as $\mathcal{S}'(\mathbb{Z}^d)$, and is defined as the set of continuous linear functionals from $\mathcal{S}(\mathbb{Z}^d)$ to $\mathbb{R}$. The Fourier transform defines a bijection $\mathcal{F} : \mathcal{C}^\infty(\mathbb{T}^d) \to S(\mathbb{Z}^d)$, with inverse

$$\mathcal{F}^{-1}a = \sum_{\xi \in \mathbb{Z}^d} a_\xi e^{2\pi i\langle \xi, \cdot \rangle}, \quad \text{for all } a \in \mathcal{S}(\mathbb{Z}^d_*)$$

which extends uniquely to a map $\mathcal{F} : \mathcal{D}'(\mathbb{T}^d) \to \mathcal{S}'(\mathbb{Z}^d)$ via the action

$$\langle \mathcal{F}u, \phi \rangle = \langle u, (\mathcal{F}\phi) \circ \iota \rangle,$$

for any test function $\phi \in \mathcal{D}(\mathbb{T}^d)$, where $\iota(x) = -x$. We represent any periodic distribution $u \in \mathcal{D}'(\mathbb{T}^d)$ by its formal power series

$$u = \sum_{\xi \in \mathbb{Z}^d} \mathcal{F}u(\xi) e^{2\pi i \langle \xi, \cdot \rangle},$$

which coincides with the classical Fourier series of u when it is sufficiently regular.

Define the Riesz kernel of order $s \in \mathbb{R}$ as the periodic distribution

$$I_s = \sum_{\xi \in \mathbb{Z}^d_*} \|\xi\|^s e^{2\pi i \langle \xi, \cdot \rangle},$$

which is in fact an $L^1(\mathbb{T}^d)$ function when $-d < s < 0$ (Stein and Weiss, 1971, Theorem 2.17), and define the fractional Laplacian of order s as the convolution operator

$$(-\Delta)^{s/2} u = I_s \star u = \mathcal{F}^{-1}\big[\| \cdot \|^s \mathcal{F}u\big], \tag{A.1}$$

for any periodic distribution $u \in \mathcal{D}'_0(\mathbb{T}^d)$. We then define the inhomogeneous Sobolev space $H^{s,r}(\mathbb{T}^d)$, for all $s \in \mathbb{R}$ and $1 < r < \infty$, as the set of periodic distributions $u \in \mathcal{D}'(\mathbb{T}^d)$ for which the norm

$$\|u\|_{H^{s,r}(\mathbb{T}^d)} = \big\|(-\Delta)^{s/2} u\big\|_{L^r(\mathbb{T}^d)}$$

is finite. Likewise, the homogeneous Sobolev space $H^{s,r}_0(\mathbb{T}^d)$ is the set of periodic distributions $u \in \mathcal{D}'_0(\mathbb{T}^d)$ such that the above norm is finite. In the special case $r = 2$, we omit the second superscript in the preceding definitions, and simply write $H^s(\mathbb{T}^d) := H^{s,2}(\mathbb{T}^d)$, $H^s_0(\mathbb{T}^d) := H^{s,2}_0(\mathbb{T}^d)$, and $\| \cdot \|_{H^s(\mathbb{T}^d)} := \| \cdot \|_{H^{s,2}(\mathbb{T}^d)}$.

Given $s \geq 0$ and $r > 1$, it follows from Theorem 3.5.6 of Schmeisser and Triebel (1987) that the $H^{-s,r}(\mathbb{T}^d)$ is isomorphic to the dual of the Banach space $H^{s,r'}(\mathbb{T}^d)$, where r' denotes the Hölder conjugate of r. Combining this fact with a similar argument as in paragraph 3.13 of Adams and Fournier (2003), one may deduce the following.

Lemma 97. *Let $r > 1$ and $s \geq 0$. Then, for all $u \in L^r_0(\mathbb{T}^d)$,*

$$\|u\|_{H^{-s,r}(\mathbb{T}^d)} \asymp \sup\Big\{ \langle u, v \rangle_{L^2(\mathbb{T}^d)} : v \in H^{s,r'}(\mathbb{T}^d), \|v\|_{H^{s,r'}(\mathbb{T}^d)} = 1 \Big\}.$$

Next, we state a standard interpolation identity. Given two complex Banach spaces B_1, B_2, we denote their θ-complex interpolation space, for any $\theta \in (0,1)$, by $(B_1, B_2)_{[\theta]}$ (Bergh and Löfström, 1976). The following can be found, for instance, in Theorem 3.6.1/2 of Schmeisser and Triebel (1987).

Lemma 98. *Let $r > 1$, $-\infty < s_0 < s_1 < \infty$, $\theta \in (0,1)$. If $s = (1-\theta)s_0 + \theta s_1$, then*

$$H^{s,r}(\mathbb{T}^d) = \big(H^{s_0,r}(\mathbb{T}^d), H^{s_1,r}(\mathbb{T}^d)\big)_{[\theta]}, \tag{A.2}$$

It will be convenient to note the following norm equivalence.

Lemma 99. *For all* $r > 1$ *and* $u \in H^{1,r}(\mathbb{T}^d)$,

$$\|u\|_{H^{1,r}(\mathbb{T}^d)} \asymp \|u\|_{L^r(\mathbb{T}^d)} + \|\nabla u\|_{L^r(\mathbb{T}^d)},$$

and for all $u \in H_0^{1,r}(\mathbb{T}^d)$,

$$\|u\|_{H^{1,r}(\mathbb{T}^d)} \asymp \|\nabla u\|_{L^r(\mathbb{T}^d)},$$

where the implicit constants in the preceding two displays depend only on d, r.

The first assertion can be deduced from Theorem 3.5.4 of Schmeisser and Triebel (1987), and the second is then a consequence of the periodic Poincaré inequality.

We will frequently make use of the following Sobolev embedding, for which a self-contained proof over the torus can be found in Bényi and Oh (2013).

Lemma 100. *Let* $s > 0$ *and* $1 < r < t < \infty$ *satisfy*

$$\frac{s}{d} = \frac{1}{r} - \frac{1}{t}.$$

Then, $\|u\|_{L^t(\mathbb{T}^d)} \lesssim \|u\|_{H^{s,r}(\mathbb{T}^d)}$ *for all* $u \in H_0^{s,r}(\mathbb{T}^d)$.

A.2.2 Sobolev Spaces over Domains

We now briefly mention a generalization of the preceding spaces to domains of $\mathbb{R}^d$, and we refer to Triebel (1995) for further details. We define the Bessel Sobolev norm of smoothness $s \in \mathbb{R}$ and integrability $1 < r < \infty$ as follows, for any tempered distribution ϕ over $\mathbb{R}^d$,

$$\|\phi\|_{H^{s,r}(\mathbb{R}^d)} = \left\|\mathcal{F}^{-1}\left[\langle\cdot\rangle^s \mathcal{F}[\phi](\cdot)\right]\right\|_{L^r(\mathbb{R}^d)},$$

and we let $H^{s,r}(\mathbb{R}^d)$ denote the completion of $\mathcal{C}_c^\infty(\mathbb{R}^d)$ under the above norm. In the special case $r = 2$, it follows from Parseval's identity that

$$\|\phi\|_{H^{s,r}(\mathbb{R}^d)} = \left\|\langle\cdot\rangle^s \mathcal{F}[\phi](\cdot)\right\|_{L^2(\mathbb{R}^d)},$$

and in this case, we omit the superscript and simply write $H^s(\mathbb{R}^d) = H^{s,2}(\mathbb{R}^d)$. Furthermore, for any domain Ω with $\mathcal{C}^\infty$ boundary, we define

$$\|\phi\|_{H^{s,r}(\Omega)} = \inf_{\substack{f \in H^{s,r}(\mathbb{R}^d) \\ \phi = f|_\Omega}} \|f\|_{H^{s,r}(\mathbb{R}^d)},$$

where the restriction $f|_\Omega$ is to be understood in the sense of distributions when $s < 0$. The space $H^{s,r}(\Omega)$ is then defined as the set of all restrictions $f|_\Omega$ of tempered distributions $f \in H^{s,r}(\Omega)$ for which the above norm is finite. Once again, we simply write $H^s(\Omega) := H^{s,2}(\Omega)$.

Finally, we define the following homogeneous Sobolev seminorms for all $s \in \mathbb{R}$,

$$\|\phi\|^2_{H^s(\mathbb{R}^d)} = \left\|\|\cdot\|^s \mathcal{F}\phi(\cdot)\right\|^2_{L^2(\mathbb{R}^d)},$$

for any $\phi \in H^s(\mathbb{R}^d)$ such that the above display exists and is finite. We define $\|\phi\|_{\dot{H}^s(\Omega)}$ analogously.

A.3 Wavelets and Besov Spaces

In Section 5.2.3 and Appendix 5.G, we make use of the boundary-corrected wavelet system Ψ^{bc} over the unit cube $[0,1]^d$, and of the periodic wavelet system Ψ^{per} over the flat torus $\mathbb{T}^d$. In this section, we provide further descriptions and properties of these wavelet bases, before turning to definitions and characterizations of Besov spaces over $[0,1]^d$ and $\mathbb{T}^d$. For concreteness, we describe these constructions in terms of the compactly-supported N-th Daubechies scaling and wavelet functions $\zeta_0, \xi_0 \in \mathcal{C}^r(\mathbb{R}^d)$, where $r = 0.18(N-1)$ for an integer $N \geq 2$ (Daubechies (1988); Giné and Nickl (2016), Theorem 4.2.10). We also extend this definition to the case $N = 1$ by taking ζ_0, ξ_0 to be the (discontinuous) Haar functions (Giné and Nickl (2016), p. 298). Throughout the thesis, whenever we work with a Besov space $\mathcal{B}^s_{p,q}$, we tacitly assume that the parameter N is chosen such that the regularity r is strictly greater than the parameters $\lceil s \rceil$ or α, in which case it must at least hold that $N \geq 2$.

Our exposition closely follows that of Giné and Nickl (2016), and we also refer the reader to Cohen, Daubechies, and Vial (1993); Cohen (2003); Härdle et al. (2012) and references therein for further details.

A.3.1 Boundary-Corrected Wavelets on $[0, 1]^d$

It is well-known that the N-th Daubechies wavelet system

$$\zeta_{0k} = \zeta_0(\cdot - k), \quad \xi_{0jk} = 2^{\frac{j}{2}}\xi_0(2^j(\cdot) - k), \quad j \geq 0,\ k \in \mathbb{Z},$$

forms a basis of $L^2(\mathbb{R})$, with the property that $\{\zeta_{0k} : k \in \mathbb{Z}\}$ spans all polynomials on $\mathbb{R}$ of degree at most $N-1$. While this family may easily be periodized to obtain a basis for $L^2([0,1])$, as in the following subsection, doing so may not accurately reflect the regularity of functions in $L^2([0,1])$ via the decay of their wavelet coefficients, near the boundaries of the interval. This consideration motivated Meyer (1991) and Cohen, Daubechies, and Vial (1993) to introduce the so-called boundary-corrected wavelet system on $[0,1]$, which preserves the standard Daubechies scaling functions lying sufficiently far from the boundaries of the interval, and adds edge scaling functions such that their union continues to span all polynomials up to degree $N-1$ on $[0,1]$. In short, given a fixed integer $j_0 \geq \log_2 N$, the construction of Cohen, Daubechies, and Vial (1993) leads to smooth scaling edge basis functions

$$\begin{aligned} &\zeta^{\mathrm{left}}_{0j_0k} &&\text{with support contained in } [0, (2N-1)/2^{j_0}], \\ &\zeta^{\mathrm{right}}_{0j_0k} &&\text{with support contained in } [1-(2N-1)/2^{j_0}, 1], \end{aligned}$$

which in turn can be used to define edge wavelet functions $\xi^{\mathrm{left}}_{0j_0k}, \xi^{\mathrm{right}}_{0j_0k}$, for $k = 0, \dots, N-1$. In this case, if one defines,

$$\zeta^a_{0jk} = 2^{\frac{j-j_0}{2}}\zeta^a_{0j_0k}\left(2^{j-j_0}(\cdot)\right), \quad \xi^a_{0jk} = 2^{\frac{j-j_0}{2}}\xi^a_{0j_0k}\left(2^{j-j_0}(\cdot)\right), \quad \text{for all } j \geq j_0,\ a \in \{\text{left, right}\},$$

then the family

$$\Phi^{\mathrm{bc}}_0 = \{\zeta^{\mathrm{bc}}_{0j_0k} : 0 \leq k \leq 2^{j_0}-1\} = \left\{\zeta^{\mathrm{left}}_{0j_0k}, \zeta^{\mathrm{right}}_{0j_0k}, \zeta_{0m} : 0 \leq k \leq N-1, N \leq m \leq 2^{j_0}-N-1\right\},$$

$$\begin{aligned}\Psi_0^{\mathrm{bc}} =& \{\xi_{0jk}^{\mathrm{bc}} : 0 \leq k \leq 2^j - 1, j \geq j_0\} \\ &= \left\{\xi_{0jk}^{\mathrm{left}}, \xi_{0jk}^{\mathrm{right}}, \xi_{0jm} : 0 \leq k \leq N-1, N \leq m \leq 2^{j_0} - N - 1, j \geq j_0\right\},\end{aligned}$$

form an orthonormal basis of $L^2([0,1])$, with the property that Φ^{bc} spans all polynomials on $[0,1]$ of degree at most $N-1$. We then define a tensor product wavelet basis of $L^2([0,1]^d)$ by setting for all $j \geq j_0$ and all $\ell = (\ell_1, \dots, \ell_d) \in \{0,1\}^d \setminus \{0\}$,

$$\zeta_{j_0 k}^{\mathrm{bc}}(x) = \prod_{i=1}^d \zeta_{0 j_0 k_i}^{\mathrm{bc}}(x_i), \quad \text{and} \quad \xi_{jk\ell}^{\mathrm{bc}}(x) = \prod_{i:\ell_i = 0} \zeta_{0jk_i}^{\mathrm{bc}}(x_i) \prod_{i:\ell_i=1} \xi_{0jk_i}^{\mathrm{bc}}(x_i), \quad x \in [0,1]^d,$$

where in the definition of $\zeta_{j_0k}^{\mathrm{bc}}$, the index $k = (k_1, \dots, k_d)$ ranges over $\mathcal{K}(j_0) := \{1, \dots, 2^{j_0} - 1\}^d$, while in the definition of $\xi_{jk\ell}^{\mathrm{bc}}$, k ranges over $\mathcal{K}(j)$. In this case, the wavelet system

$$\Psi^{\mathrm{bc}} = \Phi^{\mathrm{bc}} \cup \bigcup_{j=j_0}^{\infty} \Psi_j^{\mathrm{bc}}, \quad \Phi^{\mathrm{bc}} = \{\zeta_{j_0k}^{\mathrm{bc}} : k \in \mathcal{K}(j_0)\}, \quad \Psi_j^{\mathrm{bc}} = \{\xi_{jk\ell}^{\mathrm{bc}} : k \in \mathcal{K}(j)\}, \quad j \geq j_0,$$

announced in Section 5.2.3 forms a basis of $L^2([0,1]^d)$. We sometimes make use of the abbreviation $\Psi_{j_0-1} = \Phi$.

A.3.2 Periodic Wavelets on $\mathbb{T}^d$

When working over $\mathbb{T}^d$, a simpler construction may be used due to the periodicity of the functions involved. Denote the periodization on $\mathbb{T}$ of dilations of the maps ζ_0, ξ_0 by

$$\zeta_0^{\mathrm{per}} = \sum_{k \in \mathbb{Z}} \zeta_0(\cdot - k) = 1, \quad \xi_{0j}^{\mathrm{per}} = \sum_{k \in \mathbb{Z}} 2^{j/2} \xi_0(2^j(\cdot - k)), \quad j \geq 0.$$

In this case, the collection

$$\Psi_0^{\mathrm{per}} = \left\{1, \xi_{0jk}^{\mathrm{per}} = \xi_{0j}^{\mathrm{per}}(\cdot - 2^{-j}k) : 0 \leq k \leq 2^j - 1, j \geq 0\right\}$$

forms an orthonormal basis of $L^2(\mathbb{T})$, which may again be extended to $L^2(\mathbb{T}^d)$ using tensor product wavelets. Specifically, if $\xi_{jk\ell}^{\mathrm{per}} = \prod_{i=1}^d (\xi_{jk}^{\mathrm{per}})^{\ell_i}$ for all $\ell = (\ell_1, \dots, \ell_d) \in \{0,1\}^d \setminus \{0\}$, then

$$\Psi^{\mathrm{per}} = \{1\} \cup \bigcup_{j=0}^{\infty} \Psi_j^{\mathrm{per}}, \quad \text{with} \quad \Psi_j^{\mathrm{per}} = \left\{\xi_{jk\ell} : k \in \mathcal{K}(j), \ell \in \{0,1\}^d \setminus \{0\}\right\},\ j \geq 0,$$

forms an orthonormal basis of $L^2(\mathbb{T}^d)$ (Daubechies (1992), Section 9.3; Giné and Nickl (2016), Section 4.3).

A.3.3 Properties of Boundary-Corrected and Periodic Wavelet Systems

In both of the preceding constructions, one obtains a family Φ of scaling functions and a sequence of families $(\Psi_j)_{j\geq j_0}$ of wavelet functions, such that

$$\begin{aligned}
\Phi = \Psi_{j_0-1} = \{\zeta_k : k \in \mathcal{K}(j_0)\} &= \begin{cases} \Phi^{\mathrm{bc}}, & \text{for } \Psi = \Psi^{\mathrm{bc}} \\ \{1\}, & \text{for } \Psi = \Psi^{\mathrm{per}}, \end{cases} \\
\Psi_j = \{\xi_{jk\ell} : k \in \mathcal{K}(j), \ell \in \{0,1\}^d \setminus \{0\}\} &= \begin{cases} \Psi_j^{\mathrm{bc}}, & \text{for } \Psi = \Psi^{\mathrm{bc}} \\ \Psi_j^{\mathrm{per}}, & \text{for } \Psi = \Psi^{\mathrm{per}}, \end{cases} \qquad j \geq j_0, \\
j_0 &= \begin{cases} \lceil \log_2 N \rceil, & \text{for } \Psi = \Psi^{\mathrm{bc}} \\ 0, & \text{for } \Psi = \Psi^{\mathrm{per}}, \end{cases} \\
\mathcal{K}(j) &= \{0, \dots, 2^j - 1\}^d, \quad j \geq j_0.
\end{aligned}$$

In both cases, the wavelet system is defined over a domain Ω, which is to be understood as either $[0,1]^d$ in the boundary-corrected case, or as $\mathbb{T}^d$ (which itself may be identified with $(0,1]^d$) in the periodic case. In either of these settings, the wavelet system

$$\Psi = \Phi \cup \bigcup_{j=j_0}^{\infty} \Psi_j \tag{A.3}$$

forms a basis of $L^2(\Omega)$. The following simple result collects several properties and definitions which are common to both of the above bases.

Lemma 101. *Let $N \geq 1$. There exist constants $C_1, C_2 \geq 1$ depending only on N, d and on the choice of basis $\Psi \in \{\Psi^{\mathrm{bc}}, \Psi^{\mathrm{per}}\}$ such that the following properties hold.*

(i) The cardinalities of Φ and Ψ_j satisfy $|\Phi| \leq C_1$, $|\Psi_j| \leq C_2 2^{dj}$ for all $j \geq j_0$.

(ii) For all $j \geq j_0$ and $\xi \in \Psi_j$, there exists a rectangle $I_\xi \subseteq \Omega$ such that $\mathrm{diam}(I_\xi) \leq C_1 2^{-j}$, $\mathrm{supp}(\xi_j) \subseteq I_\xi$, and $\left\| \sum_{\xi\in\Psi_j} I(\cdot \in I_\xi) \right\|_{L^\infty} \leq C_2$.

(iii) Every element $\xi \in \Psi$ is contained in $\mathcal{C}^r(\Omega)$.

(iv) Polynomials of degree at most $N-1$ over Ω lie in $\mathrm{Span}(\Phi)$.

(v) If $N \geq 2$, we have,

$$\sup_{0\leq|\gamma|\leq\lfloor r\rfloor} \sup_{\zeta\in\Phi} \|D^\gamma \zeta\|_{L^\infty} \leq C_1, \qquad \sup_{0\leq|\gamma|\leq\lfloor r\rfloor} \sup_{j\geq j_0} \sup_{\xi\in\Psi_j} 2^{-j\left(\frac{d}{2}+|\gamma|\right)} \|D^\gamma \xi\|_{L^\infty} \leq C_2.$$

Notice that the only $\mathbb{Z}^d$-periodic polynomials on $\mathbb{R}^d$ are constants, thus Lemma 30(iv) is nearly vacuous for the basis Ψ^{per}.

A.3.4 Besov Spaces

We next define the Besov spaces $\mathcal{B}^s_{p,q}(\Omega)$, for $s > 0$, $p, q \geq 1$. Once again, Ω is understood to be one of $[0,1]^d$ or $\mathbb{T}^d$, and Ψ is understood to be the corresponding wavelet basis as in equation (A.3). Let $f \in L^p(\Omega)$ admit the wavelet expansion

$$f = \sum_{\zeta \in \Phi} \beta_\zeta \zeta + \sum_{j=j_0}^{\infty} \sum_{\xi \in \Psi_j} \beta_\xi \xi, \quad \text{over } \Omega,$$

with convergence in $L^p(\Omega)$, where $\beta_\xi = \int \xi f$ for all $\xi \in \Psi$. Then, the Besov norm of f may be defined by

$$\|f\|_{\mathcal{B}^s_{p,q}(\Omega)} = \|(\beta_\zeta)_{\zeta \in \Phi}\|_{\ell_p} + \left\| \left(2^{j(s+\frac{d}{2}-\frac{d}{p})} \left\|(\beta_\xi)_{\xi \in \Psi_j}\right\|_{\ell_p} \right)_{j \geq j_0} \right\|_{\ell_q}, \tag{A.4}$$

and we define

$$\mathcal{B}^s_{p,q}(\Omega) = \begin{cases} \left\{ f \in L^p(\Omega) : \|f\|_{\mathcal{B}^s_{p,q}(\Omega)} < \infty \right\}, & 1 \leq p < \infty \\ \left\{ f \in \mathcal{C}_u(\Omega) : \|f\|_{\mathcal{B}^s_{p,q}(\Omega)} < \infty \right\}, & p = \infty. \end{cases}$$

We extend the above definition to $s < 0$ by the duality $\mathcal{B}^s_{p',q'}(\Omega) = \left(\mathcal{B}^{-s}_{p,q}(\Omega)\right)^*$, where $\frac{1}{p'} + \frac{1}{p} = \frac{1}{q'} + \frac{1}{q} = 1$, for $p, q \notin \{1, \infty\}$. It can be shown that the resulting norm on the space $\mathcal{B}^s_{p',q'}(\Omega)$ is equivalent to the sequence norm $\|\cdot\|_{\mathcal{B}^s_{p',q'}(\Omega)}$ in equation (A.4) (cf. Cohen (2003), Theorem 3.8.1), thus we extend its definition to $s < 0$.

We will freq The following result summarizes some elementary identities (cf. Theorem 1.122 of Triebel (2006) and Section 4.3.6 of Giné and Nickl (2016)).

We shall often make use of Besov spaces in order to characterize Hölder continuous functions in terms of the decay of their wavelet coefficients, via the following classical result.

Lemma 102. *For all* $0 < s < r$*, and* $d \geq 1$*, we have*

$$\mathcal{C}^s([0,1]^d) \subseteq \mathcal{B}^s_{\infty,\infty}([0,1]^d), \quad \mathcal{C}^s(\mathbb{T}^d) \subseteq \mathcal{B}^s_{\infty,\infty}(\mathbb{T}^d), \tag{A.5}$$

and there exist $C_1, C_2 > 0$ *such that*

$$\|\cdot\|_{\mathcal{B}^s_{\infty,\infty}([0,1]^d)} \leq C_1 \|\cdot\|_{\mathcal{C}^s([0,1]^d)}, \quad \|\cdot\|_{\mathcal{B}^s_{\infty,\infty}(\mathbb{T}^d)} \leq C_2 \|\cdot\|_{\mathcal{C}^s(\mathbb{T}^d)}.$$

If $s \notin \mathbb{N}$*, then equation* (A.5) *holds with equalities, and with equivalent norms.*

An analogue of Lemma 102 is well-known to hold for the Daubechies wavelet system over $\mathbb{R}^d$, in which case it can readily be proven using an equivalent characterization of Besov spaces in terms of moduli of smoothness (Giné and Nickl (2016), Section 4.3.1). Such characterizations are also available for the periodized and boundary-corrected wavelet systems (Giné and Nickl (2016), Theorem 4.3.26 and discussions in Sections 4.3.5–4.3.6), and at least in the periodized case can be shown to lead to Lemma 102 (Giné and Nickl (2016), equation (4.167)). For the boundary-corrected case, Lemma 102 is known to hold in the special case $d = 1$ (Cohen, Daubechies, and Vial (1993), Theorem 4; Giné and Nickl (2016), equation (4.152)), but we do not know of a reference stating this precise result when $d > 1$, in part due to the potential ambiguity of defining the Hölder space $\mathcal{C}^s([0,1]^d)$ over the closed set $[0,1]^d$. We thus provide a self-contained proof of Lemma 102 in the boundary-corrected case for completeness, using standard arguments.

Proof of Lemma 102 (Boundary-Corrected Case). Let $\Omega = [0,1]^d$. Suppose first that $f \in \mathcal{B}^s_{\infty,\infty}(\Omega)$ for some $s \notin \mathbb{N}$, with wavelet expansion

$$f = \sum_{\zeta \in \Phi^{\mathrm{bc}}} \beta_\zeta \zeta + \sum_{j=j_0}^{\infty} \sum_{\xi \in \Psi_j^{\mathrm{bc}}} \beta_\xi \xi.$$

We wish to show that $\|f\|_{\mathcal{C}^s(\Omega)} \lesssim \|f\|_{\mathcal{B}^s_{\infty,\infty}(\Omega)}$. By Lemma 101, $\xi \in \mathcal{C}^r(\Omega)$ for all $\xi \in \Psi^{\mathrm{bc}}$, where recall that $\lceil s \rceil < r$, thus we may define the map

$$f_\gamma = \sum_{\zeta \in \Phi^{\mathrm{bc}}} \beta_\zeta D^\gamma \zeta + \sum_{j=j_0}^{\infty} \sum_{\xi \in \Psi_j^{\mathrm{bc}}} \beta_\xi D^\gamma \xi, \quad \text{for all } 0 \le |\gamma| \le \lfloor s \rfloor.$$

Notice that $\|D^\gamma \zeta\|_{L^\infty} \lesssim 1$ for all $\zeta \in \Phi^{\mathrm{bc}}$, and for all $j \ge j_0, k \in \mathcal{K}(j), \ell \in \{0,1\}^d \setminus \{0\}$,

$$D^\gamma \xi^{\mathrm{bc}}_{jk\ell} = 2^{(j-j_0)\left(\frac{d}{2}+|\gamma|\right)} D^\gamma \xi^{\mathrm{bc}}_{j_0 k\ell}(2^{j-j_0}(\cdot))$$

Then, it follows from Lemma 101 that for all $x \in \Omega^\circ$,

$$\begin{aligned}
|f_\gamma(x)| &\le \sum_{\zeta \in \Phi^{\mathrm{bc}}} |\beta_\zeta D^\gamma \zeta(x)| + \sum_{j=j_0}^{\infty} \sum_{\xi \in \Psi_j^{\mathrm{bc}}} |\beta_\xi D^\gamma \xi(x)| \\
&\lesssim \left\|(\beta_\zeta)_{\zeta \in \Phi^{\mathrm{bc}}}\right\|_{\ell_\infty} |\Phi^{\mathrm{bc}}| + \sum_{j=j_0}^{\infty} \|(\beta_\xi)_{\xi \in \Psi_j^{\mathrm{bc}}}\|_{\ell_\infty} 2^{(j-j_0)\left(\frac{d}{2}+|\gamma|\right)} \sum_{\xi \in \Psi_j^{\mathrm{bc}}} I(|\xi(x)| > 0) \\
&\lesssim \left\|(\beta_\zeta)_{\zeta \in \Phi^{\mathrm{bc}}}\right\|_{\ell_\infty} + \sum_{j=j_0}^{\infty} 2^{j\left(\frac{d}{2}+|\gamma|\right)} \|(\beta_\xi)_{\xi \in \Psi_j^{\mathrm{bc}}}\|_{\ell_\infty} \\
&\lesssim \left\|(\beta_\zeta)_{\zeta \in \Phi^{\mathrm{bc}}}\right\|_{\ell_\infty} + \left\|\left(2^{j\left(\frac{d}{2}+s\right)} \|(\beta_\xi)_{\xi \in \Psi_j^{\mathrm{bc}}}\|_{\ell_\infty}\right)_{j \ge j_0}\right\|_{\ell_\infty} \sum_{j=j_0}^{\infty} 2^{\frac{(|\gamma|-s)j}{2}} \lesssim \|f\|_{\mathcal{B}^s_{\infty,\infty}(\Omega)},
\end{aligned} \tag{A.6}$$

where on the final line, we used the fact that s is not an integer, thus $|\gamma| < s$. An analogous calculation reveals that the series defining f_γ converges uniformly for any $0 \le |\gamma| \le \lfloor s \rfloor$, thus it must follow that f is differentiable up to order $\lfloor s \rfloor$ with derivatives given by $D^\gamma f = f_\gamma$, which by equation (A.6) must satisfy $|D^\gamma f(x)| \le C \|f\|_{\mathcal{B}^s_{\infty,\infty}(\Omega)}$ for all $x \in \Omega^\circ$, for a constant

$C > 0$ depending only on d and r We next prove that $D^\gamma f$ is uniformly $(s - \lfloor s \rfloor)$-Hölder continuous over Ω°, for all $|\gamma| = \lfloor s \rfloor$. For all $x, y \in \Omega^\circ$, we have,

$$|D^\gamma f(x) - D^\gamma f(y)| \leq \sum_{\zeta \in \Phi^{\mathrm{bc}}} |\beta_\zeta||D^\gamma \zeta(x) - D^\gamma \zeta(y)| + \sum_{j=j_0}^{\infty} \sum_{\xi \in \Psi_j^{\mathrm{bc}}} |\beta_\xi||D^\gamma \xi(x) - D^\gamma \xi(y)|.$$

Since $\zeta \in \mathcal{C}^r(\Omega)$, for all $\zeta \in \Phi^{\mathrm{bc}}$,

$$\sum_{\zeta \in \Phi^{\mathrm{bc}}} |\beta_\zeta||D^\gamma \zeta(x) - D^\gamma \zeta(y)| \lesssim \|f\|_{\mathcal{B}^s_{\infty,\infty}(\Omega)} |\Phi^{\mathrm{bc}}| \|x - y\| \lesssim \|f\|_{\mathcal{B}^s_{\infty,\infty}(\Omega)} \|x - y\| .$$

Furthermore, using the definition of the boundary-corrected wavelet basis and its locality property in Lemma 30(ii), we have

$$\begin{aligned}
&\sum_{j=j_0}^{\infty} \sum_{\xi \in \Psi_j^{\mathrm{bc}}} |\beta_\xi||D^\gamma \xi(x) - D^\gamma \xi(y)| \\
&\quad = \sum_{j=j_0}^{\infty} \sum_{k=0}^{2^j-1} \sum_{l \in \{0,1]^d \setminus \{0\}} |\beta_{\xi_{jk\ell}}| 2^{(j-j_0)\left(\frac{d}{2}+|\gamma|\right)} |D^\gamma \xi_{j_0 k\ell}(2^{j-j_0}(x)) - D^\gamma \xi_{j_0 k\ell}(2^{j-j_0}(y))| \\
&\quad \lesssim \sum_{j=j_0}^{\infty} \|(\beta_\xi)_{\xi \in \Psi_j^{\mathrm{bc}}}\|_{\ell_\infty} 2^{(j-j_0)\left(\frac{d}{2}+|\gamma|\right)} \left(\|2^{j-j_0}x - 2^{j-j_0}y\| \wedge 1\right) \sum_{\xi \in \Psi_j^{\mathrm{bc}}} I(|\xi(x)| \vee |\xi(y)| > 0) \\
&\quad \lesssim \sum_{j=j_0}^{\infty} \|(\beta_\xi)_{\xi \in \Psi_j^{\mathrm{bc}}}\|_{\ell_\infty} 2^{(j-j_0)\left(\frac{d}{2}+|\gamma|\right)} \left(2^{j-j_0}\|x - y\| \wedge 1\right) \\
&\quad \lesssim \sum_{j=0}^{\infty} \|(\beta_\xi)_{\xi \in \Psi_{j+j_0}^{\mathrm{bc}}}\|_{\ell_\infty} 2^{j\left(\frac{d}{2}+|\gamma|\right)} \left(2^j\|x - y\| \wedge 1\right) \\
&\quad \lesssim \sum_{j=0}^{J(x,y)} \|(\beta_\xi)_{\xi \in \Psi_{j+j_0}^{\mathrm{bc}}}\|_{\ell_\infty} 2^{j\left(\frac{d}{2}+|\gamma|+1\right)} \|x - y\| + \sum_{j=J(x,y)}^{\infty} \|(\beta_\xi)_{\xi \in \Psi_{j+j_0}^{\mathrm{bc}}}\|_{\ell_\infty} 2^{j\left(\frac{d}{2}+|\gamma|\right)},
\end{aligned}$$

where $J(x, y)$ is the smallest integer $j \geq 0$ such that $2^j|x - y| \geq 1$; in particular,

$$2^{-J(x,y)} \leq \|x - y\| \leq 2^{-J(x,y)+1}. \tag{A.7}$$

Now, since $2^{j(\frac{d}{2}+s)}\|(\beta_\xi)_{\xi \in \Psi_{j+j_0}^{\mathrm{bc}}}\|_{\ell_\infty} \leq \|f\|_{\mathcal{B}^s_{\infty,\infty}(\Omega)} < \infty$, and since $|\gamma| < s \notin \mathbb{N}$, we obtain

$$\begin{aligned}
\|f\|^{-1}_{\mathcal{B}^s_{\infty,\infty}(\Omega)} |D^\gamma f(x) - D^\gamma f(y)| &\lesssim \|x - y\| \sum_{j=0}^{J(x,y)} 2^{j(|\gamma|-s+1)} + \sum_{j=J(x,y)}^{\infty} 2^{j(|\gamma|-s)} \\
&\lesssim \|x - y\| 2^{J(x,y)(|\gamma|-s+1)} + 2^{J(x,y)(|\gamma|-s)} \lesssim \|x - y\|^{s-|\gamma|} ,
\end{aligned}$$

where the final inequality is due to equation (A.7). It readily follows that $\|f\|_{\mathcal{C}^s(\Omega)} \lesssim \|f\|_{\mathcal{B}^s_{\infty,\infty}(\Omega)}$. Furthermore, since $D^\gamma f$ is uniformly Hölder continuous over $(0,1)^d$, it is in particular uniformly continuous and hence extends to a continuous function over $[0,1]^d$, thus $f \in \mathcal{C}^s([0,1]^d)$. We next show that $\mathcal{C}^s([0,1]^d) \subseteq \mathcal{B}^s_{\infty,\infty}([0,1]^d)$ for all $s > 0$, with the requisite Hölder norms. Assume $\|f\|_{\mathcal{C}^s(\Omega)} < \infty$, and let $\beta_\xi = \int f\xi$ for all $\xi \in \Psi^{\mathrm{bc}}$. By definition of the Besov norm, it will suffice to prove that

$$\left\|(\beta_\zeta)_{\zeta\in\Phi^{\mathrm{bc}}}\right\| \lesssim \|f\|_{\mathcal{C}^s([0,1]^d)}, \quad \left\|(\beta_\xi)_{\xi\in\Psi_j^{\mathrm{bc}}}\right\| \lesssim \|f\|_{\mathcal{C}^s([0,1]^d)}\, 2^{-j(\frac{d}{2}+s)}, \quad j \geq j_0.$$

The first bound is immediate, since f is bounded above by $\|f\|_{\mathcal{C}^s([0,1]^d)}$ over $[0,1]^d$. To prove the second bound, let $x_0 \in (0,1)^d$, and let $\underline{s}$ denote the largest integer strictly less than s. By a Taylor expansion to order $\underline{s}$, there exists $c_s > 0$ such that

$$\left| f(x) - \sum_{0\leq|\gamma|\leq\underline{s}} D^\gamma f(x_0)(x-x_0)^\gamma \right| \leq c_s \|f\|_{\mathcal{C}^s(\Omega)} \|x-x_0\|^s, \quad x \in \Omega, \tag{A.8}$$

where $(x-x_0)^\gamma = \prod_{i=1}^d (x_i - x_{0i})^{\gamma_i}$. In particular, for any given $\xi \in \Psi_j^{\mathrm{bc}}$, $j \geq j_0$, choose $x_0 \in I_\xi \cap (0,1)^d$, where $\mathrm{diam}(I_\xi) \lesssim 2^{-j}$ and I_ξ is a set containing the support of ξ, as defined in Lemma 30(ii). We then have,

$$\begin{aligned}\left|\int \xi f\right| &\lesssim \left| \int \xi(x) \sum_{0\leq|\gamma|\leq\underline{s}} D^\gamma f(x_0)(x-x_0)^\gamma dx \right| \\ &\quad + \|f\|_{\mathcal{C}^s(\Omega)} \int |\xi(x)|\, \|x-x_0\|^s\, dx = \|f\|_{\mathcal{C}^s(\Omega)} \int |\xi(x)|\, \|x-x_0\|^s\, dx,\end{aligned}$$

where the final equality uses the fact that polynomials of degree at most $\lfloor r \rfloor$ lie in $\mathrm{Span}(\Phi^{\mathrm{bc}})$ by Lemma 30(iv), and are therefore orthogonal to ξ. We thus have,

$$\begin{aligned}|\beta_\xi| &\lesssim \|f\|_{\mathcal{C}^s(\Omega)} \int_\Omega |\xi(x)|\, \|x-x_0\|^s\, dx \\ &= \|f\|_{\mathcal{C}^s(\Omega)} \int_{I_\xi} |\xi(x)|\, \|x-x_0\|^s\, dx \\ &\lesssim \|f\|_{\mathcal{C}^s(\Omega)} 2^{dj/2}\, \mathrm{diam}(I_\xi)^s \mathcal{L}(I_\xi) \lesssim \|f\|_{\mathcal{C}^s(\Omega)} 2^{-j(\frac{d}{2}+s)}.\end{aligned}$$

The claim readily follows. □

Appendix B

Wavelet and Kernel Density Estimation

In this Appendix, we state several properties of wavelet and kernel density estimators which are used throughout our development. We always work over $[0,1]^d$ or $\mathbb{T}^d$ in what follows.

B.1 Wavelet Density Estimation

We begin by discussing wavelet density estimators over $\Omega \in \{\mathbb{T}^d, [0,1]^d\}$, with the corresponding basis $\Psi \in \{\Psi^{\mathrm{per}}, \Psi^{\mathrm{bc}}\}$ as in Section A.3.3. Let $q \in L^2(\Omega)$ denote a probability density with corresponding probability distribution Q, and with corresponding wavelet expansion

$$q = \sum_{\zeta\in\Phi} \beta_\zeta \zeta + \sum_{j=j_0}^{\infty} \sum_{\xi\in\Psi_j} \beta_\xi \xi.$$

Given an i.i.d. sample $Y_1, \dots, Y_n \sim Q$ with corresponding empirical measure $Q_n = (1/n)\sum_{i=1}^n \delta_{Y_i}$, define the unnormalized and normalized wavelet density estimators of the density q of Q,

$$\widetilde{q}_n = \sum_{\zeta\in\Phi} \widehat{\beta}_\zeta \zeta + \sum_{j=j_0}^{J_n} \sum_{\xi\in\Psi_j} \widehat{\beta}_\xi \xi, \qquad \widehat{q}_n = \frac{\widetilde{q}_n I(\widetilde{q}_n \geq 0)}{\int_{\widetilde{q}_n\geq 0} \widetilde{q}_n d\mathcal{L}}, \tag{B.1}$$

where $J_n \geq j_0$ is a deterministic threshold, and $\widehat{\beta}_\xi = \int \xi dQ_n$ for all $\xi \in \Psi_j$, $j_0 \leq j \leq J_n$. The following simple result guarantees that $\widetilde{q}_n$ integrates to unity since q is a probability density.

Lemma 103. *We have $\int_\Omega \widetilde{q}_n = 1$. In particular, it follows that $\sum_{\zeta\in\Phi} \widehat{\beta}_\zeta \int_\Omega \zeta = 1$.*

The proof of Lemma 103 appears in Appendix B.1.0.1. In the special case of the periodic wavelet system, for which Φ^{per} consists only of the constant function 1, Lemma 103 implies that the corresponding estimated coefficient satisfies $\widehat{\beta}_1 = 1$ deterministically, thus the definition of $\widetilde{q}_n$ in equation (B.1) coincides with that which will be given in Appendix 5.G.

With this result in place, we turn to L^∞ concentration results for $\widetilde{q}_n$, as well as for Besov norms of $\widetilde{q}_n$, which we frequently use throughout our proofs. In what follows, write

$$q_{J_n}(y) = \mathbb{E}[\widetilde{q}_{J_n}(y)] = \sum_{\zeta \in \Phi} \beta_\zeta \zeta + \sum_{j=j_0}^{J_n} \sum_{\xi \in \Psi_j} \beta_\xi \xi, \quad y \in \Omega.$$

Lemma 104. *Let $N \geq 2$ and $q \in \mathcal{B}^s_{\infty,\infty}(\Omega)$ for some $s > 0$. Then, there exist constants $v, b > 0$ depending only on the choice of wavelet system, such that for any $J_n \geq j_0$, and all $u > 0$,*

$$\mathbb{P}\left(\sup_{\zeta \in \Phi} |\widehat{\beta}_\zeta - \beta_\zeta| \geq u \right) \lesssim \exp\left\{ -\frac{nu^2}{b} \right\}, \tag{B.2}$$

$$\mathbb{P}\left(\sup_{\xi \in \Psi_j} |\widehat{\beta}_\xi - \beta_\xi| \geq u \right) \lesssim 2^{\frac{dj}{2}} \exp\left\{ -\frac{nu^2}{v + 2^{jd/2} bu} \right\}, \quad j_0 \leq j \leq J_n. \tag{B.3}$$

Furthermore, if $2^{J_n} = c_0 n^{\frac{1}{d+2s}}$ for some $c_0 > 0$, then there exists a constant $C > 0$, depending on c_0 and on the choice of wavelet system Ψ, such that the following assertions hold.

(i) For all $0 < u \leq 1$,

$$\mathbb{P}\Big(\|\widetilde{q}_n\|_{\mathcal{B}^{s/2}_{\infty,\infty}(\Omega)} \geq u + \|q\|_{\mathcal{B}^{s/2}_{\infty,\infty}(\Omega)} \Big) \leq C J_n 2^{dJ_n} \exp\big(-u^2 2^{sJ_n}/C \big).$$

(ii) For all $2^{-J_n} \leq u \leq 1$,

$$\mathbb{P}\left(\|\widetilde{q}_n - q_{J_n}\|_{L^\infty(\Omega)} \geq u \right) \leq C J_n 2^{J_n d(d+3)} \exp\big(-nu^2 2^{-dJ_n}/C \big).$$

Lemma 30(ii) is implicit in the proofs of almost sure L^∞ bounds for wavelet estimators by Masry (1997) and Guo and Kou (2019), as well as Giné and Nickl (2009) when $d = 1$. While these results are based on wavelet estimators over $\mathbb{R}^d$, they can readily be adapted to the wavelet systems considered here, as consequences of inequalities (B.2)–(B.3). For completeness, we provide a proof of Lemma 30(ii), along with the remaining assertions of Lemma 104, in Appendix B.1.0.2.

Using Lemmas 103 and 30(ii), the following result is now straightforward.

Lemma 105. *Let $N \geq 2$. Assume there exist $\gamma, s > 0$ such that $q \geq 1/\gamma$ over Ω, and such that $q \in \mathcal{B}^s_{\infty,\infty}(\Omega)$. Then, there exists $c_1 > 0$ depending on γ, $\|q\|_{\mathcal{B}^s_{\infty,\infty}(\Omega)}$ such that with probability at least $1 - c_1/n^2$, $\widetilde{q}_n$ is a valid probability density over Ω, and hence $\widehat{q}_n = \widetilde{q}_n$. If we instead have $N = 1$, then under no conditions on q it holds that $\widehat{q}_n = \widetilde{q}_n$ almost surely.*

Having now established that $\widetilde{q}_n$ is a valid density with high probability, we may speak of its convergence in Wasserstein distance. Niles-Weed and Berthet (2022) previously derived upper bounds on the risk, in Wasserstein distance over $[0,1]^d$, of a projection of $\widetilde{q}_n$ onto the set of probability densities. Using Lemma 105, we are able to extend their result to the estimator $\widehat{Q}_n$, i.e. the distribution function of the density $\widehat{q}_n$ defined in equation (B.1). We also state this result for a general exponent of the 2-Wasserstein distance.

Lemma 106. *Let $\Psi = \Psi^{\mathrm{bc}}$ with $N \geq 2$. Assume that $q \in \mathcal{B}^s_{\infty,\infty}([0,1]^d)$ for some $s > 0$. Assume further that $q \geq 1/\gamma$ over $[0,1]^d$ for some $\gamma > 0$. Let $2^{J_n} \asymp n^{1/(d+2s)}$. Then, for any $\rho \geq 0$, there exists a constant $C > 0$ depending on M, γ, ρ, s such that*

$$\mathbb{E}W_2^\rho(\widehat{Q}_n, Q) \leq C \begin{cases} n^{-\frac{\rho(s+1)}{2s+d}}, & d \geq 3 \\ (\log n/\sqrt{n})^\rho, & d = 2 \\ 1/n^{\rho/2}, & d = 1. \end{cases} \tag{B.4}$$

Furthermore, when $N = 1$, equation (B.4) *continues to hold with $s = 0$ for any density satisfying $\gamma^{-1} \leq q \leq \gamma$ over $[0,1]^d$, for some $\gamma > 0$.*

The proof appears in Appendix B.1.0.4.

B.1.0.1 Proof of Lemma 103

Recall that $\mathrm{Span}(\Phi)$ contains all polynomials of degree at most $N-1$ over Ω, by Lemma 30(iv). In particular, it contains the constant function 1, thus if $\beta'_\zeta = \int_\Omega \zeta$, we obtain $1 = \sum_{\zeta\in\Phi} \beta'_\zeta \zeta$. It then follows by orthonormality of Ψ that

$$\int_\Omega \widetilde{q}_n = \int_\Omega \left(\sum_{\zeta\in\Phi} \beta'_\zeta \zeta\right)\left(\sum_{\zeta\in\Phi} \widehat{\beta}_\zeta \zeta + \sum_{j=j_0}^{J_n} \sum_{\xi\in\Psi_j} \widehat{\beta}_\xi \xi\right) = \sum_{\zeta\in\Phi} \beta'_\zeta \widehat{\beta}_\zeta = \int \left(\sum_{\zeta\in\Phi} \beta'_\zeta \zeta\right) dQ_n = 1.$$

This proves the claim. □

B.1.0.2 Proof of Lemma 104

Throughout the proof, $b, v, c > 0$ denote constants depending only on c_0 and the choice of wavelet system, whose value may change from line to line. To prove inequality (B.2), recall first from Lemma 30(v) that

$$\sup_{\zeta\in\Phi} \|\zeta\|_{L^\infty(\Omega)} \leq b, \qquad \sup_{j\geq j_0} 2^{-jd/2} \sup_{\xi\in\Psi_j} \|\xi\|_{L^\infty(\Omega)} \leq b. \tag{B.5}$$

By Hoeffding's inequality, equation (B.5) implies that for all $u > 0$,

$$\mathbb{P}\left(\sup_{\zeta\in\Phi} |\widehat{\beta}_\zeta - \beta_\zeta| \geq u\right) \leq \sum_{\zeta\in\Phi} \mathbb{P}\left(\left|\int \zeta d(Q_n - Q)\right| \geq u\right) \lesssim \exp\left\{-\frac{nu^2}{b^2}\right\}, \tag{B.6}$$

where we have used the fact that $|\Phi| \lesssim 1$ by Lemma 30(i). To prove equation (B.3), notice that for all $\xi \in \Psi_j$ and $j \geq j_0$, given $Y \sim Q$,

$$\mathrm{Var}[\xi(Y)] \leq \int \xi^2(y)q(y)dy \leq \|q\|_{L^\infty(\Omega)} \int \xi^2(y)dy = \|q\|_{L^\infty(\Omega)} \leq v,$$

where we used the fact that $q \in \mathcal{B}^s_{\infty,\infty}(\Omega) \subseteq L^\infty(\Omega)$. Therefore, by Bernstein's inequality, we have for all $u > 0$ and $j_0 \leq j \leq J_n$,

$$\mathbb{P}\left(\sup_{\xi\in\Psi_j} |\widehat{\beta}_\xi - \beta_\xi| \geq u\right) \leq \sum_{\xi\in\Psi_j} \mathbb{P}\Big(|\widehat{\beta}_\xi - \beta_\xi| \geq u\Big) \lesssim 2^{dj} \exp\left\{-\frac{nu^2}{v + 2^{jd/2}bu}\right\}. \tag{B.7}$$

Here, the last inequality uses the fact that $|\Psi_j| \lesssim 2^{dj}$ by Lemma 30(i) for all $j \geq j_0$.

To prove part (i) from here, let $0 < u \leq 1$. A union bound combined with the above display leads to

$$\mathbb{P}\left(\sup_{j_0\leq j\leq J_n} \sup_{\xi\in\Psi_j} |\widehat{\beta}_\xi - \beta_\xi| \geq u\right) \lesssim J_n 2^{dJ_n} \exp\left\{-\frac{nu^2}{v + 2^{J_n d/2}bu}\right\}, \tag{B.8}$$

whence, since $2^{J_n} \asymp n^{\frac{1}{d+2s}}$,

$$\begin{aligned}&\mathbb{P}\left(2^{\frac{J_n(s+d)}{2}} \sup_{j_0\leq j\leq J_n} \left\|(\widehat{\beta}_\xi - \beta_\xi)_{\xi\in\Psi_j}\right\|_{\ell_\infty} \geq u\right) \\ &\qquad\lesssim J_n 2^{dJ_n} \exp\left\{-\frac{nu^2 2^{-J_n(s+d)}}{v{+}b2^{\frac{dJ_n}{2}}2^{-\frac{J_n s}{2}-\frac{dJ_n}{2}}u}\right\} \leq J_n 2^{dJ_n} \exp\left\{-cu^2 2^{J_n s}\right\}.\end{aligned} \tag{B.9}$$

Combining this fact with equation (B.6), we have

$$\begin{aligned}&\mathbb{P}\left(\|\widetilde{q}_n - q_{J_n}\|_{\mathcal{B}^{s/2}_{\infty,\infty}} \geq u\right) \\ &\leq \mathbb{P}\left(\left\|(\widehat{\beta}_\zeta - \beta_\zeta)_{\zeta\in\Phi}\right\|_{\ell_\infty} \geq u/2\right) + \mathbb{P}\left(2^{\frac{J_n(d+s)}{2}} \sup_{j_0\leq j\leq J_n} \left\|(\widehat{\beta}_\xi - \beta_\xi)_{\xi\in\Psi_j}\right\|_{\ell_\infty} \geq u/2\right) \\ &\leq CJ_n 2^{dJ_n} \exp\{-u^2 2^{J_n s}/C\},\end{aligned}$$

for a large enough constant $C > 0$. Thus, we have

$$\|\widetilde{q}_n\|_{\mathcal{B}^{s/2}_{\infty,\infty}} \leq \|\widetilde{q}_n - q_{J_n}\|_{\mathcal{B}^{s/2}_{\infty,\infty}} + \|q_{J_n}\|_{\mathcal{B}^{s/2}_{\infty,\infty}} \leq u + \|q\|_{\mathcal{B}^{s/2}_{\infty,\infty}}$$

with probability at least $1 - CJ_n 2^{dJ_n} \exp\{-u^2 2^{J_n s}/C\}$. Part (i) thus follows. To prove part (ii), let $\delta_n \leq 2^{J_n(d+2)}/(4C_0)$, for a constant $C_0 > 0$ to be specified below. Notice that for all $x, y \in \Omega$,

$$\begin{aligned}|\widetilde{q}_n(x) - \widetilde{q}_n(y)| &\leq \left|\sum_{\zeta\in\Phi} \widehat{\beta}_\zeta(\zeta(x) - \zeta(y))\right| + \left|\sum_{j=j_0}^{J_n} \sum_{\xi\in\Psi_j} \widehat{\beta}_\xi(\xi(x) - \xi(y))\right| \\ &\lesssim \sum_{\zeta\in\Phi} |\widehat{\beta}_\zeta| \|x - y\| + \sum_{j=j_0}^{J_n} 2^{j\left(\frac{d}{2}+1\right)} |\widehat{\beta}_\xi| \|x - y\| \sum_{\xi\in\Psi_j} I(\xi(x) \wedge \xi(y) > 0)\end{aligned}$$

$$
\begin{aligned}
&\lesssim \|x-y\| + \sum_{j=j_0}^{J_n} 2^{j\left(\frac{d}{2}+1\right)} \|\xi\|_{L^\infty(\Omega)} \|x-y\| \\
&\lesssim \sum_{j=j_0}^{J_n} 2^{j(d+1)} \|x-y\| \lesssim 2^{J_n(d+1)} \|x-y\|,
\end{aligned}
$$

where we have again used the properties appearing in Lemma 101. Upon repeating an analogous calculation, we deduce that both $\widetilde{q}_n$ and q_{J_n} are $C_0 2^{J_n(d+1)}$-Lipschitz.

Let $K_n = O(1/\delta_n^d) = O(2^{-J_n d(d+2)})$ denote the δ_n-covering number of the unit cube $[0,1]^d$ with respect to the Euclidean norm, and let $\{x_{0k} : 1 \leq k \leq K_n\}$ be a corresponding δ_n-cover. Letting $I_k = \{x \in [0,1]^d : \|x - x_{0k}\| \leq \delta_n\}$, we have (for both $\Omega \in \{[0,1]^d, \mathbb{T}^d\}$),

$$
\begin{aligned}
\|\widetilde{q}_n - q_{J_n}\|_{L^\infty(\Omega)} &\leq \max_{1\leq k\leq K_n} \sup_{x\in I_k} |\widetilde{q}_n(x) - q_{J_n}(x)| \\
&\leq \max_{1\leq k\leq K_n} \sup_{x\in I_k} |\widetilde{q}_n(x) - \widetilde{q}_n(x_{0k})| \\
&\qquad + \max_{1\leq k\leq K_n} \sup_{x\in I_k} |q_{J_n}(x_{0k}) - q_{J_n}(x)| + \max_{1\leq k\leq K_n} |\widetilde{q}_n(x_{0k}) - q_{J_n}(x_{0k})| \\
&\leq 2C_0 2^{J_n(d+1)}\delta_n + \max_{1\leq k\leq K_n} |\widetilde{q}_n(x_{0k}) - q_{J_n}(x_{0k})| \\
&\leq 2^{-J_n}/2 + \max_{1\leq k\leq K_n} |\widetilde{q}_n(x_{0k}) - q_{J_n}(x_{0k})|.
\end{aligned}
$$

Thus, for any $2^{-J_n} \leq u \leq 1$, using Lemma 101 and the bounds (B.6)–(B.8), we have

$$
\begin{aligned}
&\mathbb{P}\left(\|\widetilde{q}_n - q_{J_n}\|_{L^\infty(\Omega)} \geq u\right) \\
&\quad\leq \mathbb{P}\left(\max_{1\leq k\leq K_n} |\widetilde{q}_n(x_{0k}) - q_{J_n}(x_{0k})| \geq u/2\right) \\
&\quad\leq \sum_{k=1}^{K_n} \mathbb{P}\left(\left|\sum_{\zeta\in\Phi} (\widehat{\beta}_\zeta - \beta_\zeta)\zeta(x_{0k}) + \sum_{j=j_0}^{J_n} \sum_{\xi\in\Psi_j} (\widehat{\beta}_\xi - \beta_\xi)\xi(x_{0k})\right| \geq u/2\right) \\
&\quad\leq \sum_{k=1}^{K_n} \mathbb{P}\left(\left|\sum_{\zeta\in\Phi} (\widehat{\beta}_\zeta - \beta_\zeta)\zeta(x_{0k})\right| \geq u/4\right) + \sum_{k=1}^{K_n} \mathbb{P}\left(\left|\sum_{j=j_0}^{J_n} \sum_{\xi\in\Psi_j} (\widehat{\beta}_\xi - \beta_\xi)\xi(x_{0k})\right| \geq u/4\right) \\
&\quad\leq K_n \mathbb{P}\left(\sup_{\zeta\in\Phi} |\widehat{\beta}_\zeta - \beta_\zeta| \geq cu\right) + K_n \mathbb{P}\left(J_n 2^{\frac{dJ_n}{2}} \sup_{j_0\leq j\leq J_n} \sup_{\xi\in\Psi_j} |\widehat{\beta}_\xi - \beta_\xi| \geq cu\right) \\
&\quad\lesssim K_n \exp(-nc^2u^2/b^2) + J_n K_n 2^{dJ_n} \exp\left(-nc^2u^2 2^{-dJ_n}/(J_n^2 v + cbJ_n u)\right).
\end{aligned}
$$

It follows that, for a sufficiently large constant $C > 0$,

$$
\mathbb{P}\left(\|\widetilde{q}_n - q_{J_n}\|_{L^\infty(\Omega)} \geq u\right) \leq CJ_n 2^{J_n d(d+3)} \exp\left(- nu^2 2^{-dJ_n}/(J_n C)\right),
$$

for all $2^{-J_n} < u \leq 1$. The claim readily follows. □

B.1.0.3 Proof of Lemma 105

The claim for $N = 1$ follows by definition of the Haar system, since in this case $\widetilde{q}_n$ is equal to a histogram. We thus assume $s > 0$ and $N \geq 2$. Recall that $\widetilde{q}_n$ integrates to unity by Lemma 103, thus it suffices to show that $\widetilde{q}_n \geq 0$ with high probability. Apply Lemma 104 to deduce that

$$\|\widetilde{q}_n - q_{J_n}\|_{L^\infty(\Omega)} \leq \gamma^{-1}/4,$$

except on an event with probability at most c_1/n^2, for some $c_1 > 0$ depending on γ^{-1} and $\|q\|_{\mathcal{B}^s_{\infty,\infty}(\Omega)}$. Furthermore, using Lemma 101, the bias of $\widetilde{q}_n$ satisfies

$$\begin{aligned}\|q_{J_n} - q\|_{L^\infty(\Omega)} &= \sum_{j \geq J_n+1} 2^{\frac{dj}{2}} \|(\beta_\xi)_{\xi \in \Psi_j}\|_{\ell_\infty} \\ &\leq \|q\|_{\mathcal{B}^s_{\infty,\infty}(\Omega)} \sum_{j \geq J_n+1} 2^{\frac{dj}{2} - j(\frac{d}{2}+s)} \lesssim \|q\|_{\mathcal{B}^s_{\infty,\infty}(\Omega)} 2^{-J_n s} \leq \gamma^{-1}/4,\end{aligned}$$

for all n larger than a universal constant depending only on $\|q\|_{\mathcal{B}^s_{\infty,\infty}(\Omega)}$. Therefore, after possibly increasing $c_1 > 0$, we have with probability at least $1 - c_1/n^2$ that for all $n \geq 1$,

$$\|\widetilde{q}_n - q\|_{L^\infty(\Omega)} \leq \gamma^{-1}/2.$$

Since $q \geq \gamma^{-1}$, we deduce that $\widetilde{q}_n \geq \gamma^{-1}/2 \geq 0$, over the same high probability event. □

B.1.0.4 Proof of Lemma 106

By Jensen's inequality, it suffices to assume that $\rho \geq 1$. It is straightforward to verify from Lemma 101 that the wavelet system Ψ^{bc} satisfies Assumptions E.1–E.6 of Niles-Weed and Berthet (2022), except Assumption E.2 in the special case $N = 1$. We also have $\gamma^{-1} \leq q \leq \gamma$ over $[0,1]^d$. These conditions are sufficient to invoke their Theorem 4 for any $N \geq 1$, leading to

$$W_2(\widehat{Q}_n, Q) \lesssim_\gamma \|\widehat{q}_n - q\|_{\mathcal{B}^{-1}_{2,1}([0,1]^d)}\,.$$

Furthermore, it follows from Lemma 105 that the event $A_n = \{\widehat{q}_n = \widetilde{q}_n\}$ satisfies $\mathbb{P}(A_n^{\mathsf{c}}) \lesssim n^{-2}$. Let $q_{J_n} = \mathbb{E}[\widetilde{q}_n]$, so that

$$\mathbb{E}W_2^\rho(\widehat{Q}_n, Q) = \mathbb{E}\Big[W_2^\rho(\widehat{Q}_n, Q) I_{A_n}\Big] + \mathbb{E}\Big[W_2^\rho(\widehat{Q}_n, Q) I_{A_n^{\mathsf{c}}}\Big] \lesssim \mathbb{E}\|\widetilde{q}_n - q\|^\rho_{\mathcal{B}^{-1}_{2,1}([0,1]^d)} + n^{-2}.$$

Now, we make use of the following result which can be deduced from the proofs of Theorem 1 and Proposition 4 of Niles-Weed and Berthet (2022).

Lemma 107 (Niles-Weed and Berthet (2022)). *Let q be a density satisfying $\gamma^{-1} \leq q \leq \gamma$ over $[0,1]^d$. Assume further that $q \in \mathcal{B}^s_{\infty,\infty}([0,1]^d)$ for some $s \geq 0$. Then,*

$$\begin{aligned}&\|q_{J_n} - q\|^\rho_{\mathcal{B}^{-1}_{2,1}([0,1]^d)} \lesssim 2^{-\rho J_n(s+1)}, \\ &\mathbb{E}\big\|(\widehat{\beta}_\zeta - \beta_\zeta)_{\zeta \in \Phi^{\mathrm{bc}}}\big\|^\rho_{\ell_2} \lesssim 1/n^{\rho/2}, \quad \mathbb{E}\big\|(\widehat{\beta}_\xi - \beta_\xi)_{\xi \in \Psi^{\mathrm{bc}}_j}\big\|^\rho_{\ell_2} \lesssim \Big(2^{dj}/n^{1/2}\Big)^\rho, \quad j \geq j_0.\end{aligned}$$

Let $\rho' \geq 1$ satisfy $\frac{1}{\rho} + \frac{1}{\rho'} = 1$. Lemma 107 implies,

$$\begin{aligned}
&\mathbb{E}\|\widetilde{q}_n - q\|^{\rho}_{\mathcal{B}^{-1}_{2,1}([0,1]^d)} \\
&\quad\lesssim \mathbb{E}\|\widetilde{q}_n - q_{J_n}\|^{\rho}_{\mathcal{B}^{-1}_{2,1}([0,1]^d)} + \|q_{J_n} - q\|^{\rho}_{\mathcal{B}^{-1}_{2,1}([0,1]^d)} \\
&\quad\lesssim \mathbb{E}\big\|(\widehat{\beta}_\zeta - \beta_\zeta)_{\zeta\in\Phi^{\mathrm{bc}}}\big\|^{\rho}_{\ell_2} + \mathbb{E}\left(\sum_{j=j_0}^{J_n} 2^{-j}\big\|(\widehat{\beta}_\xi - \beta_\xi)_{\xi\in\Psi_j^{\mathrm{bc}}}\big\|_{\ell_2}\right)^{\rho} + 2^{-\rho J_n(s+1)} \\
&\quad\lesssim n^{-\frac{\rho}{2}} + \left(\sum_{j=j_0}^{J_n} 2^{\rho(\eta-1)j}\mathbb{E}\big\|(\widehat{\beta}_\xi - \beta_\xi)_{\xi\in\Psi_j^{\mathrm{bc}}}\big\|^{\rho}_{\ell_2}\right)\left(\sum_{j=j_0}^{J_n} 2^{-\rho'\eta j}\right)^{\frac{\rho}{\rho'}} + 2^{-\rho J_n(s+1)} \\
&\quad\lesssim n^{-\frac{\rho}{2}} + n^{-\frac{\rho}{2}}\left(\sum_{j=j_0}^{J_n} 2^{\rho(\eta+\frac{d}{2}-1)j}\right)\left(\sum_{j=j_0}^{J_n} 2^{-\rho'\eta j}\right)^{\frac{\rho}{\rho'}} + 2^{-\rho J_n(s+1)},
\end{aligned}$$

for any $\eta \in \mathbb{R}$. Now, when $d \geq 3$, choose $1 - \frac{d}{2} < \eta < 0$. In this case, the above display is of order

$$n^{-\frac{\rho}{2}} 2^{[\rho(\eta+\frac{d}{2}-1)-\rho\eta]J_n} + 2^{-\rho J_n(s+1)} = 2^{\rho(\frac{d}{2}-1)J_n} + 2^{-\rho J_n(s+1)} \lesssim n^{-\frac{\rho(s+1)}{2s+d}},$$

which proves the claim for $d \geq 3$. When $d \leq 2$, choose $\eta = 0$. Then, the penultimate display is dominated by its second term, which is of order $n^{-\rho/2}$ when $d = 1$ and of order $(\log n/\sqrt{n})^{\rho}$ when $d = 2$. The claim follows. □

B.2 Kernel Density Estimation

Recall that we define the kernel density estimator

$$\widetilde{q}_n = K_{h_n} \star Q_n = \int_{\mathbb{R}^d} K_{h_n}(\cdot - y)dQ_n(y),$$

where, in the above display, $Q_n \in \mathcal{P}(\mathbb{T}^d)$ is extended by $\mathbb{Z}^d$-periodicity to a Borel measure on $\mathbb{R}^d$. Under condition **K($\boldsymbol{\alpha}$)**, we assume that $K \in \mathcal{C}_c^\infty(0,1)^d$, and thus it is easy to see that the periodization of K_{h_n}, denoted by

$$\overline{K}_{h_n} = \sum_{\xi\in\mathbb{Z}^d} K_{h_n}(\cdot - \xi),$$

defines a function in $\mathcal{C}^\infty(\mathbb{T}^d)$, with the property that for all $x \in \mathbb{R}^d$, there exists $\xi \in \mathbb{Z}^d$ such that $K_{h_n}(x) = \overline{K}_{h_n}(x - \xi)$. By the Poisson summation formula, it holds that

$$\mathcal{F}[\overline{K}_{h_n}](\xi) = \mathcal{F}[K_{h_n}](\xi) = \mathcal{F}[K](h_n\xi), \quad \text{for all } \xi \in \mathbb{Z}^d. \tag{B.10}$$

Throughout what follows, we write for all $u \in L^1(\mathbb{T}^d)$,

$$\mathcal{K}_{h_n}u = \overline{K}_{h_n} \star u - u.$$

The aim of this appendix is to derive the convergence rate of the kernel density estimator under the Hölder norms $\mathcal{C}^\gamma(\mathbb{T}^d)$ with $\gamma \geq 0$, and under the negative Sobolev norms $H^{-\gamma,r}(\mathbb{T}^d)$, $r \geq 2$. We begin with the former.

B.2.1 Convergence Rate under Hölder Norms

When $\gamma \geq 0$ is an integer, the question of characterizing the convergence rate of $\widetilde{q}_n$ under the $\mathcal{C}^\gamma(\mathbb{T}^d)$ norm reduces to deriving the uniform convergence rate of derivatives of the kernel density estimator, which is a classical topic (Bhattacharya, 1967; Silverman, 1978; Giné and Nickl, 2016). Using an elementary interpolation argument, we can extend this existing work to non-integer values of γ, as stated next.

Proposition 42. Let $0 \leq \gamma < s$, and assume $q \in \mathcal{C}^s(\mathbb{T}^d)$. Let K satisfy condition **K(s)**. Then, there exist constants $C, b > 0$ depending on $\|q\|_{\mathcal{C}^s(\mathbb{T}^d)}, \gamma, s$ such that for any $h_n \geq 0$,

$$\mathbb{E}\|q_{h_n} - q\|_{\mathcal{C}^\gamma(\mathbb{T}^d)} \leq C h_n^{s-\gamma},$$

and such that for any $h_n \leq u \leq 1$,

$$\mathbb{P}\Big(\|\widetilde{q}_n - q_{h_n}\|_{\mathcal{C}^\gamma(\mathbb{T}^d)} \geq u\Big) \leq C h_n^{-b} \exp\big(-nu^2 h_n^{2\gamma+d}/C\big).$$

In particular, there exists a constant $C_1 > 0$, depending in particular on q, such that, almost surely for all $n \geq 1$,

$$\|\widetilde{q}_n - q_{h_n}\|_{\mathcal{C}^\gamma(\mathbb{T}^d)} \leq C_1 \sqrt{\frac{\log(h_n^{-1})}{n h_n^{2\gamma+d}}}.$$

Proof of Proposition 42. Given $a, b \geq 0$ and a bounded linear operator $F : \mathcal{C}^a(\mathbb{T}^d) \to \mathcal{C}^b(\mathbb{T}^d)$, denote the norm of F by

$$\|F\|_{\mathcal{C}^a(\mathbb{T}^d)\to\mathcal{C}^b(\mathbb{T}^d)} = \sup_{u \in \mathcal{C}^a(\mathbb{T}^d)\setminus\{0\}} \frac{\|Fu\|_{\mathcal{C}^a(\mathbb{T}^d)}}{\|u\|_{\mathcal{C}^b(\mathbb{T}^d)}}.$$

When γ is an integer, it is a standard fact that the following bound holds under condition **K(s)** (see for instance Giné and Nickl (2016), page 402):

$$\|\mathcal{K}_{h_n}\|_{\mathcal{C}^s(\mathbb{T}^d)\to\mathcal{C}^\gamma(\mathbb{T}^d)} \lesssim h_n^{s-\gamma}.$$

If instead $\gamma > 0$ is not an integer, let $\gamma_0 = \lfloor \gamma \rfloor$, $\gamma_1 = \lceil \gamma \rceil$ and $\theta = \gamma/\gamma_1$. The periodic Hölder space $\mathcal{C}^\gamma(\mathbb{T}^d)$ is a real interpolation space of exponent θ between $\mathcal{C}^{\gamma_0}(\mathbb{T}^d)$ and $\mathcal{C}^{\gamma_1}(\mathbb{T}^d)$ (cf. Schmeisser and Triebel (1987), page 173), whence it follows that

$$\begin{aligned}\|\mathcal{K}_{h_n}\|_{\mathcal{C}^s(\mathbb{T}^d)\to\mathcal{C}^\gamma(\mathbb{T}^d)} &\lesssim \|\mathcal{K}_{h_n}\|^{1-\theta}_{\mathcal{C}^s(\mathbb{T}^d)\to\mathcal{C}^{\gamma_0}(\mathbb{T}^d)} \|\mathcal{K}_{h_n}\|^{\theta}_{\mathcal{C}^s(\mathbb{T}^d)\to\mathcal{C}^{\gamma_1}(\mathbb{T}^d)} \\ &\lesssim \big(h_n^{s-\gamma_0}\big)^{1-\theta}\big(h_n^{s-\gamma_1}\big)^{\theta} = h_n^{s-\gamma}.\end{aligned}$$

This proves the first claim. To prove the second claim, it can be deduced from the proof of Lemma 32 of Manole et al. (2021) (see also Giné and Guillou (2002)) that for all integers $\gamma \geq 0$, there exists $C > 0$ such that for all $h_n < u \leq 1$,

$$\mathbb{P}\Big(\|\widetilde{q}_n - q_{h_n}\|_{\mathcal{C}^\gamma(\mathbb{T}^d)} \geq u/h_n^\gamma\Big) \leq Ch_n^{-d(d+\gamma+2)} \exp\big(-nu^2h_n^d/C\big).$$

If instead $\gamma > 0$ is a real number, it follows again from the interpolation property of the periodic Hölder spaces that the following inequality holds (cf. Lunardi (2018, Corollary 1.7)),

$$\|\widetilde{q}_n - q_{h_n}\|_{\mathcal{C}^\gamma(\mathbb{T}^d)} \lesssim \|\widetilde{q}_n - q_{h_n}\|_{\mathcal{C}^{\lfloor\gamma\rfloor}}^{1-\frac{\gamma}{\lceil\gamma\rceil}} \|\widetilde{q}_n - q_{h_n}\|_{\mathcal{C}^{\lceil\gamma\rceil}}^{\frac{\gamma}{\lceil\gamma\rceil}}.$$

The preceding two displays imply,

$$\begin{aligned}
&\mathbb{P}\left(\|\widetilde{q}_n - q_{h_n}\|_{\mathcal{C}^\gamma(\mathbb{T}^d)} \geq u/h_n^\gamma\right) \\
&\leq \mathbb{P}\left(\|\widetilde{q}_n - q_{h_n}\|_{\mathcal{C}^{\lfloor\gamma\rfloor}(\mathbb{T}^d)} \geq u/h_n^{\lfloor\gamma\rfloor}\right) + \mathbb{P}\left(\|\widetilde{q}_n - q_{h_n}\|_{\mathcal{C}^{\lceil\gamma\rceil}(\mathbb{T}^d)} \geq u/h_n^{\lceil\gamma\rceil}\right) \\
&\lesssim h_n^{-d(d+\lceil\gamma\rceil+2)} \exp\big(-nu^2h_n^d/C\big),
\end{aligned}$$

from which the second claim follows. The final claim is now a direct consequence of the first Borel-Cantelli Lemma. □

We deduce the following Lemma.

Lemma 108. *Let $q \in \mathcal{C}_+^s(\mathbb{T}^d)$ for some $s > 0$, and let $0 \leq \gamma < s$. Let K satisfy condition* **K(*s*)**. *Suppose that for some $c > 0$,*

$$h_n = c \cdot n^{-a}, \quad \textit{with } 0 < a < \frac{1}{2\gamma + d},$$

Then, for any $b > 0$, there exist constants $C, \delta > 0$ depending on K, γ, b, c such that with probability at least $1 - C/n^b$.

1. *$\|\widetilde{q}_n\|_{\mathcal{C}^\gamma(\mathbb{T}^d)} \leq C$.*
2. *$\widetilde{q}_n$ is a valid density; in particular, $\widetilde{q}_n \geq C^{-1} > 0$ and $\int_{\mathbb{T}^d} \widetilde{q}_n = 1$.*

In particular, there exist constants $C, N > 0$, possibly depending on q, such that the two preceding assertions hold almost surely for all $n \geq N$.

In particular, the following is an immediate consequence of Lemma 108 and Theorem 3.

Lemma 109. *Let $p, q \in \mathcal{C}_+^s(\mathbb{T}^d)$ and let $0 < \gamma < s$, $\gamma \notin \mathbb{N}$. Suppose that for some $c > 0$,*

$$h_n = c \cdot n^{-a}, \quad \textit{with } 0 < a < \frac{1}{2\gamma + d}.$$

Then, for any $b > 0$, there exist constants C, λ depending on $\omega_s(p, q), \gamma, b, c$ such that with probability at least $1 - C/n^b$, the unique mean-zero Brenier potential $\widehat{\varphi}_n$ whose gradient pushes forward p onto $\widetilde{q}_n$ is well-defined and satisfies

$$\|\widehat{\varphi}_n\|_{\mathcal{C}^{\gamma+2}(\mathbb{T}^d)} \leq \lambda, \quad \textit{and } \widehat{\varphi}_n \textit{ is } \lambda^{-1}\textit{-strongly convex.}$$

In particular, there exist constants $C, N, \lambda > 0$ depending on p and q such that the above display holds almost surely for all $n \geq N$.

B.2.2 Convergence Rate under Negative Sobolev Norms

We now turn to the question of bounding the convergence rate of the kernel density estimator under the negative Riesz potential Sobolev norms.

Proposition 43. Let $0 < \alpha < d$, $s \geq 0$, $r \geq 2$, and $q \in \mathcal{C}^s_+(\mathbb{T}^d)$. Assume K satisfies condition **K($s + \alpha$)**. Assume further that for some $c > 0$, $h_n \geq c \cdot n^{-1/d}$. Then, there exists a constant $C = C(\omega_s(q), K, c, \alpha, s, r) > 0$ such that

$$\|q_{h_n} - q\|_{H^{-\alpha,r}(\mathbb{T}^d)} \leq C h_n^{s+\alpha},$$

and,

$$\mathbb{E}\|\widetilde{q}_n - q_{h_n}\|_{H^{-\alpha,r}(\mathbb{T}^d)} \leq \Big(\mathbb{E}\|\widetilde{q}_n - q_{h_n}\|^r_{H^{-\alpha,r}(\mathbb{T}^d)}\Big)^{\frac{1}{r}} \leq C\frac{h_n^{\alpha-\frac{d}{2}}}{\sqrt{n}}.$$

Furthermore,

$$C^{-1}\frac{h_n^{2\alpha-d}}{n} \leq \mathbb{E}\|\widetilde{q}_n - q_{h_n}\|^2_{H^{-\alpha}(\mathbb{T}^d)} \leq C\frac{h_n^{2\alpha-d}}{n}.$$

When α = 0, the above result reduces to the traditional convergence rate of the kernel density estimator under $L^r(\mathbb{T}^d)$ norms (Giné and Nickl, 2016). When $s = 0$ and $\alpha = 1$, this result has previously appeared in Divol (2021), and our proof below is inspired by their approach.

Proof of Proposition 43. We begin with the bound on the bias of $\widetilde{q}_n$. Recall the definition of the fractional Laplacian $(-\Delta)^{-\alpha/2}$ and its associated Riesz kernel I_α, given in Chapter A. We have,

$$\begin{aligned}\|q_{h_n} - q\|_{H^{-\alpha,r}(\mathbb{T}^d)} &= \big\|I_\alpha \star \mathcal{K}_{h_n}[q]\big\|_{L^r(\mathbb{T}^d)} \\ &= \big\|\mathcal{K}_{h_n}[I_\alpha \star q]\big\|_{L^r(\mathbb{T}^d)} \\ &= \big\|\mathcal{K}_{h_n}\big[(-\Delta)^{-\alpha/2} q\big]\big\|_{L^r(\mathbb{T}^d)}. \end{aligned} \tag{B.11}$$

Now, let $\gamma = \alpha + s$, and notice that by definition of the Riesz potential spaces,

$$\big\|(-\Delta)^{-\alpha/2} q\big\|_{H^{\gamma,r}(\mathbb{T}^d)} \lesssim \|q\|_{H^{s,r}(\mathbb{T}^d)}. \tag{B.12}$$

Therefore, the question of bounding the expression in equation (B.11) reduces to the question of bounding the $L^r(\mathbb{T}^d)$ approximation error of the convolution of the $H^{\gamma,r}(\mathbb{T}^d)$ function $f := (-\Delta)^{-\alpha/2}[q]$. When γ is an integer (Giné and Nickl, 2016), but we provide below a direct proof for general γ. Specifically, we aim to show

$$\|\mathcal{K}_{h_n}[f]\|_{L^r(\mathbb{T}^d)} \lesssim h_n^\gamma.$$

Define $M := \|\cdot\|^{-\gamma}\big(\mathcal{F}[K](\cdot) - 1\big)$, with $0/0 = 0$. We will show below that $M(a\cdot)$ is a $L^r(\mathbb{T}^d)$ Fourier multiplier for any given $a > 0$, in the sense that

$$\big\|\mathcal{F}^{-1}[M(a\cdot)\mathcal{F}f]\big\|_{L^r(\mathbb{T}^d)} \leq C\|f\|_{L^r(\mathbb{T}^d)}$$

for all $f \in L^r(\mathbb{T}^d)$, for a constant $C > 0$ independent of a. Taking this fact for granted momentarily, let us show how the claim follows. Using equation (B.10), we have

$$\begin{aligned}\|\mathcal{K}_{h_n}[f]\|_{L^r(\mathbb{T}^d)} &= \big\|\mathcal{F}^{-1}\big(\mathcal{F}[\mathcal{K}_{h_n}[f]\big)\big\|_{L^r(\mathbb{T}^d)} \\ &= h_n^\gamma\big\|\mathcal{F}^{-1}\big(M(h_n\xi)\|\xi\|^\gamma\mathcal{F}[f](\xi)\big)\big\|_{L^r(\mathbb{T}^d)} \\ &\lesssim h_n^\gamma\big\|\mathcal{F}^{-1}\big(\|\xi\|^\gamma\mathcal{F}[f](\xi)\big)\big\|_{L^r(\mathbb{T}^d)} = h_n^\gamma\|f\|_{H^{\gamma,r}(\mathbb{T}^d)} \lesssim h_n^{s+\alpha}\|q\|_{H^{s,r}(\mathbb{T}^d)},\end{aligned}$$

where we used equation (B.12) to obtain the final inequality.

It thus remains to show that $M(a\cdot)$ is indeed a Fourier multiplier, with norm independent of $a > 0$. To do so, it will suffice to show that M satisfies the conditions of Mikhlin's multiplier theorem (Grafakos, 2008, Theorem 6.2.7). Abbreviate $g = \mathcal{F}[K]$. By condition **K($\boldsymbol{\gamma}$)**, M is bounded over $\mathbb{R}^d$. Furthermore, recall that condition **K($\boldsymbol{\gamma}$)** implies that K is a kernel of order $\gamma - 1$, i.e. $D^\kappa g(0) = 0$ for all multi-indices κ such that $1 \leq |\kappa| \leq \gamma - 1$. For any such κ, we have

$$D^\kappa g(\xi) = D^\kappa g(0) + \sum_{\alpha:1\leq|\omega|\leq\gamma-1-|\kappa|} D^{\kappa+\omega}g(0)\xi^\omega + O(\|\xi\|^{\gamma-|\kappa|}) = O(\|\xi\|^{\gamma-|\kappa|}),$$

for $\|\xi\| \leq 1$. Since K and g are Schwartz functions, the above bound also continues to hold trivially for all $|\kappa| \geq \gamma$. Using the general Leibniz rule, we deduce that for all multi-indices ω satisfying $1 \leq |\omega| \leq \frac{d}{2} + 1$, we have[1]

$$\begin{aligned}\big|D^\omega M(\xi)\big| &= \big|D^\omega\big(g(\xi)\|\xi\|^{-\gamma}\big)\big| \\ &= \left|\sum_{\kappa\leq\omega}\binom{\omega}{\kappa}\big(D^\kappa g(\xi)\big)\big(D^{\omega-\kappa}\|\xi\|^{-\gamma}\big)\right| \\ &\lesssim \sum_{\kappa\leq\omega}\big|D^\kappa g(\xi)\big|\big|D^{\omega-\kappa}\|\xi\|^{-\gamma}\big| \\ &\lesssim \sum_{\kappa\leq\omega}\|\xi\|^{\gamma-|\kappa|}\|\xi\|^{-\gamma-|\omega-\kappa|} \\ &\lesssim \|\xi\|^{-|\omega|},\end{aligned}$$

for all $\|\xi\| \leq 1$. Finally, since g is a Schwarz function, the last bound of the above display continues to hold trivially when $\|\xi\| > 1$. The conditions of Mikhlin's Multiplier Theorem are thus satisfied. This completes the proof of the bias bound.

[1] The notation $\kappa \leq \omega$ is to be understood componentwise. Furthermore, we write $\binom{\omega}{\kappa} = \prod_{i=1}^d \binom{\omega}{\kappa_i}$.

To prove the first claim about the variance, we follow the proof of Divol (2021) closely. We would like to bound $V = \mathbb{E}\|(1/n)\sum_{i=1}^n \big(U_i - \mathbb{E}U_i\big)\|_{L^r(\mathbb{T}^d)}^r$, where

$$U_i(x) = I_\alpha \star \overline{K}_{h_n}^o(X_i - x), \quad x \in \mathbb{T}^d,$$

where we write $\overline{K}_{h_n}^o = \overline{K}_{h_n} - 1$. Using Rosenthal's inequality (Rosenthal, 1970, 1972), it holds that

$$V \lesssim n^{-r/2}\int_{\mathbb{T}^d} \big(\mathbb{E}|U_1(x)|^2\big)^{r/2} dx + n^{1-r}\int_{\mathbb{T}^d} \mathbb{E}|U_1(x)|^r dx.$$

Notice that, due to the upper-boundedness of q, one has for any $x \in \mathbb{T}^d$,

$$\begin{aligned}\mathbb{E}|U_1(x)|^r &= \int_{\mathbb{T}^d} \big|I_\alpha \star \overline{K}_{h_n}^o(y-x)\big|^r q(y)dy \\ &\lesssim \int_{\mathbb{T}^d} \big|I_\alpha \star \overline{K}_{h_n}^o(y-x)\big|^r dy = \int_{\mathbb{T}^d} \big|I_\alpha \star \overline{K}_{h_n}^o(y)\big|^r dy = \|\overline{K}_{h_n}^o\|_{H^{-\alpha,r}(\mathbb{T}^d)}^r,\end{aligned}$$

where we used the translational invariance of $\mathcal{L}$. We now make use of the following.

Lemma 110. *Let* $0 < \alpha < d$, $r \geq 2$, *and assume condition* **K($\boldsymbol{\alpha}$)** *holds. Then, there exists* $C = C(\alpha, r) > 0$ *such that*

$$C^{-1} h_n^{r\alpha-(r-1)d} \leq \|\overline{K}_{h_n}^o\|_{H^{-\alpha,r}(\mathbb{T}^d)}^r \leq C h_n^{r\alpha-(r-1)d}.$$

The proof appears below. We thus have

$$V \lesssim n^{-r/2}\big(h_n^{2\alpha-d}\big)^{r/2} + n^{1-r}h_n^{r\alpha-(r-1)d} = n^{-r/2}h_n^{r\left(\alpha-\frac{d}{2}\right)} + n^{1-r}h_n^{r\alpha-(r-1)d}. \tag{B.13}$$

Note that

$$c\cdot n^{-r/2}h_n^{r\left(\alpha-\frac{d}{2}\right)} \geq n^{1-r}h_n^{r\alpha-(r-1)d} \iff h_n \geq c\cdot n^{-1/d},$$

thus, under our assumption that $h_n \geq cn^{-1/d}$, the first term on the right-hand side of equation (B.13) dominates, and we obtain

$$V \lesssim n^{-r/2}h_n^{r\left(\alpha-\frac{d}{2}\right)}.$$

It thus remains to prove the final claim, for which it suffices to prove the lower bound. We have,

$$\begin{aligned}\mathbb{E}\|\widetilde{q}_n - q_{h_n}\|_{H^{-\alpha}(\mathbb{T}^d)}^2 &= \mathbb{E}\|(1/n)\sum_{i=1}^n \big(U_i - \mathbb{E}U_i\big)\|_{L^2(\mathbb{T}^d)}^2 \\ &\asymp \frac{1}{n}\int_{\mathbb{T}^d}\int_{\mathbb{T}^d} \big|I_\alpha \star \overline{K}_{h_n}^o(y-x)\big|^2 dxdy \\ &= \frac{1}{n}\|\overline{K}_{h_n}^o\|_{H^{-\alpha}(\mathbb{T}^d)}^2 \asymp \frac{1}{n}h_n^{2\alpha-d},\end{aligned}$$

where we used the fact that q is bounded from above and below over $\mathbb{T}^d$, and we invoked Lemma 110 in the final order assessment. The claim follows. □

Proof of Lemma 110. Let

$$\kappa = \sup_{\xi \in \mathbb{R}^d \setminus \{0\}} \frac{|\mathcal{F}[K](\xi) - 1|}{\|\xi\|},$$

which is finite by condition **K(1)**. By the Hausdorff-Young inequality, notice that

$$\|\overline{K}^o_{h_n}\|_{H^{-\alpha,r}(\mathbb{T}^d)} = \big\|\mathcal{F}^{-1}\big[\|\cdot\|^{-\alpha}\mathcal{F}[\overline{K}^o_{h_n}]\big]\big\|_{L^r(\mathbb{T}^d)} \leq \big\|\big(\|\cdot\|^{-\alpha}\mathcal{F}[\overline{K}^o_{h_n}]\big)_{\xi \in \mathbb{Z}^d_*}\big\|_{\ell^{r'}(\mathbb{Z}^d_*)},$$

where $r' = r/(r-1)$ is the Hölder conjugate of r. We thus have,

$$\|\overline{K}^o_{h_n}\|^{r'}_{H^{-\alpha,r}(\mathbb{T}^d)} \leq \sum_{\xi \in \mathbb{Z}^d_*} \frac{|\mathcal{F}[K](h_n\xi)|^{r'}}{\|\xi\|^{\alpha r'}},$$

where

$$S_{n,1} = \sum_{\substack{\xi \in \mathbb{Z}^d_* \\ \kappa\|h_n\xi\| \leq 1/2}} \frac{\big|\mathcal{F}[K](h_n\xi)\big|^{r'}}{\|2\pi\xi\|^{r'\alpha}}, \qquad S_{n,2} = \sum_{\substack{\xi \in \mathbb{Z}^d_* \\ \kappa\|h_n\xi\| > 1/2}} \frac{\big|\mathcal{F}[K](h_n\xi)\big|^{r'}}{\|2\pi\xi\|^{r'\alpha}}.$$

Regarding term $S_{n,1}$, apply condition **K($\boldsymbol{\alpha}$)** to obtain

$$\begin{aligned} S_{n,1} &\geq \sum_{\substack{\xi \in \mathbb{Z}^d_* \\ \kappa\|h_n\xi\| \leq 1/2}} \frac{(1 - \kappa\|h_n\xi\|)^{r'}_+}{\|2\pi\xi\|^{r'\alpha}} \\ &\geq \frac{1}{2^{r'}\|2\pi\|^{r'\alpha}} \sum_{\substack{\xi \in \mathbb{Z}^d_* \\ \kappa\|h_n\xi\| \leq 1/2}} \frac{1}{\|\xi\|^{r'\alpha}} \asymp h_n^{r'\alpha - d}, \end{aligned}$$

where the final order assessment holds due to the condition $r\alpha < d(r-1)$, which implies $r'\alpha < d$. A similar argument shows that $S_{n,1} \lesssim h_n^{r'\alpha - d}$. It thus remains to upper bound $S_{n,2}$. Recall that $K \in \mathcal{C}^\infty_c(\mathbb{R}^d)$, thus K and $\mathcal{F}[K]$ both belong to the Schwartz space. In particular, $\mathcal{F}[K](\xi) \lesssim \|\xi\|^{-\ell}$ for any $\ell > 0$. Choosing ℓ such that $r'(\ell + \alpha) > d$, it follows that

$$S_{n,2} \lesssim h_n^{-r'\ell} \sum_{\substack{\xi \in \mathbb{Z}^d_* \\ \kappa\|h_n\xi\|^\rho > 1/2}} \|\xi\|^{-r'(\ell+\alpha)} \lesssim h_n^{-r'\ell} h_n^{-d + r'(\ell+\alpha)} = h_n^{r'\alpha - d}.$$

Deduce from here that

$$\|\overline{K}^o_{h_n}\|^r_{H^{-\alpha,r}(\mathbb{T}^d)} \asymp h_n^{\frac{r}{r'}(r'\alpha - d)} = h_n^{r\alpha - (r-1)d}.$$

This proves the claim. □

Bibliography

Abbasi-Yadkori, Y., Pal, D., and Szepesvari, C. (2011). Online Least Squares Estimation with Self-Normalized Processes: An Application to Bandit Problems. *arXiv preprint arXiv:1102.2670.*

Adams, R. A. and Fournier, J. J. (2003). *Sobolev Spaces.* Elsevier.

Agrawal, R. (2020). Finite-sample concentration of the multinomial in relative entropy. *IEEE Transactions on Information Theory*, 66:6297–6302.

Agrawal, R. and Horel, T. (2020). Optimal bounds between f-divergences and integral probability metrics. In *Proceedings of the Thirty-Seventh International Conference on Machine Learning*, PMLR, pages 115–124.

Ajtai, M., Komlós, J., and Tusnády, G. (1984). On optimal matchings. *Combinatorica*, 4:259–264.

Ali, S. M. and Silvey, S. D. (1966). A General Class of Coefficients of Divergence of One Distribution from Another. *Journal of the Royal Statistical Society: Series B*, 28:131–142.

Alison, J. (2015). *The Road to Discovery: Detector Alignment, Electron Identification, Particle Misidentification, WW Physics, and the Discovery of the Higgs Boson.* PhD thesis, University of Pennsylvania.

Álvarez-Esteban, P. C., Del Barrio, E., Cuesta-Albertos, J. A., and Matran, C. (2008). Trimmed comparison of distributions. *Journal of the American Statistical Association*, 103:697–704.

Alwall, J., Herquet, M., Maltoni, F., Mattelaer, O., and Stelzer, T. (2011). MadGraph 5: Going beyond. *Journal of High Energy Physics*, 2011:128.

Aly, E.-E. A. (1986). Strong approximations of the QQ process. *Journal of Multivariate Analysis*, 20:114–128.

Ambrosio, L., Colombo, M., De Philippis, G., and Figalli, A. (2012). Existence of Eulerian solutions to the semigeostrophic equations in physical space: The 2-dimensional periodic case. *Communications in Partial Differential Equations*, 37:2209–2227.

Ambrosio, L., Gigli, N., and Savare, G. (2008). *Gradient Flows: In Metric Spaces and in the Space of Probability Measures.* Lectures in Mathematics. ETH Zürich. Birkhäuser Basel, 2 edition.

Ambrosio, L. and Glaudo, F. (2019). Finer estimates on the 2-dimensional matching problem. *Journal de l'École polytechnique—Mathématiques*, 6:737–765.

Ambrosio, L., Glaudo, F., and Trevisan, D. (2019). On the optimal map in the 2-dimensional random matching problem. *arXiv preprint arXiv:1903.12153*.

Ambrosio, L., Stra, F., and Trevisan, D. (2019). A PDE approach to a 2-dimensional matching problem. *Probability Theory and Related Fields*, 173:433–477.

Arjovsky, M., Chintala, S., and Bottou, L. (2017). Wasserstein generative adversarial networks. In *Proceedings of the 34th International Conference on Machine Learning*, pages 214–223.

ATLAS (2012). Observation of a new particle in the search for the Standard Model Higgs boson with the ATLAS detector at the LHC. *Physics Letters B*, 716:1–29.

ATLAS (2015). Observation and measurement of Higgs boson decays to WW^* with the ATLAS detector. *Physical Review D*, 92:012006.

ATLAS (2018a). Measurement of gluon fusion and vector-boson-fusion higgs boson production cross-sections in the $H \to WW^* \to e\nu\mu\nu$ decay channel in pp collisions at $\sqrt{s} = 13$ TeV with the ATLAS detector. ATLAS-CONF-2018-004.

ATLAS (2018b). Measurements of Higgs boson properties in the diphoton decay channel with 36 fb- 1 of p p collision data at s= 13 TeV with the ATLAS detector. *Physical Review D*, 98:052005.

ATLAS (2018c). Measurements of the higgs boson production, fiducial and differential cross sections in the 4ℓ decay channel at $\sqrt{s} = 13$ TeV with the ATLAS detector. ATLAS-CONF-2018-018.

ATLAS (2018d). Observation of $H \to b\bar{b}$ decays and vh production with the ATLAS detector. *Physics Letters B*, 786:59–86.

ATLAS (2018). Search for pair production of higgsinos in final states with at least three b-tagged jets in 13 TeV pp collisions using the atlas detector. *Physical Review D*, 98.

ATLAS (2019a). Cross-section measurements of the higgs boson decaying into a pair of τ-leptons in proton-proton collisions at $\sqrt{s} = 13$ TeV with the ATLAS detector. *Physical Review D*, 99:072001.

ATLAS (2019b). Search for pair production of Higgs bosons in the $b\bar{b}b\bar{b}$ final state using proton-proton collisions at $\sqrt{s} = 13$ TeV with the ATLAS detector. *Journal of High Energy Physics*, 2019:30.

ATLAS (2021). Search for the hh$\to$ $b\bar{b}b\bar{b}$ process via vector-boson fusion production using proton-proton collisions at $\sqrt{s} = 13$ TeV with the atlas detector. *Journal of High Energy Physics*, 2021.

ATLAS (2022). Search for non-resonant pair production of Higgs bosons in the $b\bar{b}b\bar{b}$ final state in pp collisions at $\sqrt{s} = 13$ TeV with the ATLAS detector. Technical report, CERN, Geneva. ATLAS-CONF-2022-035.

ATLAS (2023). Search for ZZ and ZH production in the $b\bar{b}b\bar{b}$ final state using proton-proton collisions at $\sqrt{s}$ = 13 TeV. *CMS-PAS-HIG-22-011*.

ATLAS, CMS, and Higgs Combination Group (2011). Procedure for the LHC Higgs boson search combination in Summer 2011. CMS-NOTE-2011-005. ATL-PHYS-PUB-2011-11.

Balsubramani, A. and Ramdas, A. (2016). Sequential nonparametric testing with the law of the iterated logarithm. In *Proceedings of the Thirty-Second Conference on Uncertainty in Artificial Intelligence*, pages 42–51.

Barlow, R. (1987). Event classification using weighting methods. *Journal of Computational Physics*, 72:202–219.

Behnke, O., Kröninger, K., Schott, G., and Schörner-Sadenius, T. (2013). *Data Analysis in High Energy Physics: A Practical Guide to Statistical Methods.* John Wiley & Sons.

Benamou, J.-D. and Brenier, Y. (2000). A computational fluid mechanics solution to the Monge-Kantorovich mass transfer problem. *Numerische Mathematik*, 84:375–393.

Bényi, Á. and Oh, T. (2013). The Sobolev inequality on the torus revisited. *Publicationes Mathematicae Debrecen*, 83:359.

Berend, D. and Kontorovich, A. (2013). A sharp estimate of the binomial mean absolute deviation with applications. *Statistics and Probability Letters*, 83:1254–1259.

Bergh, J. and Löfström, J. (1976). *Interpolation Spaces: An Introduction*, volume 223 of *Grundlehren Der Mathematischen Wissenschaften.* Springer.

Bernton, E., Jacob, P. E., Gerber, M., and Robert, C. P. (2019a). Approximate Bayesian computation with the Wasserstein distance. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 81:235–269.

Bernton, E., Jacob, P. E., Gerber, M., and Robert, C. P. (2019b). On parameter estimation with the Wasserstein distance. *Information and Inference: A Journal of the IMA*, 8:657–676.

Berrett, T. B. and Samworth, R. J. (2023). Efficient functional estimation and the super-oracle phenomenon. *The Annals of Statistics*, 51:668–690.

Berthet, P., Fort, J.-C., and Klein, T. (2020). A central limit theorem for Wasserstein type distances between two distinct univariate distributions. *Annales de l'Institut Henri Poincaré, Probabilités et Statistiques*, 56:954–982.

Bhattacharya, P. K. (1967). Estimation of a probability density function and its derivatives. *Sankhyā: The Indian Journal of Statistics, Series A*, 29:373–382.

Bickel, P. J. and Freedman, D. A. (1981). Some asymptotic theory for the bootstrap. *The Annals of Statistics*, 9:1196–1217.

Bickel, P. J. and Ritov, Y. (1988). Estimating integrated squared density derivatives: Sharp best order of convergence estimates. *Sankhyā: The Indian Journal of Statistics, Series A*, pages 381–393.

Billingsley, P. (1968). *Convergence of Probability Measures.* John Wiley & Sons.

Birgé, L. and Massart, P. (1995). Estimation of integral functionals of a density. *The Annals of Statistics*, 23:11–29.

Bladt, M. and Shaiderman, D. (2023). On the characterization of exchangeable sequences through reverse-martingale empirical distributions. *Electronic Communications in Probability*, 28:1–11.

Bobkov, S. and Ledoux, M. (2019). One-dimensional empirical measures, order statistics, and Kantorovich transport distances. *Memoirs of the American Mathematical Society*, 261.

Bobkov, S. G. and Götze, F. (1999). Exponential integrability and transportation cost related to logarithmic Sobolev inequalities. *Journal of Functional Analysis*, 163:1–28.

Bogachev, V. I. (2007). *Measure Theory.* Springer-Verlag, Berlin, Germany.

Boissard, E. and Le Gouic, T. (2014). On the mean speed of convergence of empirical and occupation measures in Wasserstein distance. In *Annales de l'Institut Henri Poincaré Probabilités et Statistiques*, volume 50, pages 539–563.

Bonneel, N., Rabin, J., Peyré, G., and Pfister, H. (2015). Sliced and Radon Wasserstein barycenters of measures. *Journal of Mathematical Imaging and Vision*, 51:22–45.

Bonneel, N., Van De Panne, M., Paris, S., and Heidrich, W. (2011). Displacement interpolation using Lagrangian mass transport. In *Proceedings of the 2011 SIGGRAPH Asia Conference*, pages 1–12.

Bonnotte, N. (2013). *Unidimensional and Evolution Methods for Optimal Transportation.* PhD thesis, Paris 11.

Borisyak, M. and Kazeev, N. (2019). Machine Learning on data with sPlot background subtraction. *Journal of Instrumentation*, 14:P08020–P08020.

Borwein, J. M. and Borwein, P. B. (1987). *Pi and the AGM: A Study in the Analytic Number Theory and Computational Complexity.* Wiley-Interscience.

Boucheron, S., Lugosi, G., and Massart, P. (2013). *Concentration Inequalities: A Nonasymptotic Theory of Independence.* Oxford University Press.

Bousquet, O., Boucheron, S., and Lugosi, G. (2003). Introduction to statistical learning theory. In *Summer School on Machine Learning*, pages 169–207. Springer.

Brasco, L., Carlier, G., and Santambrogio, F. (2010). Congested traffic dynamics, weak flows and very degenerate elliptic equations. *J. Math. Pures Appl. (9)*, 93:163–182.

Brehmer, J., Kling, F., Espejo, I., and Cranmer, K. (2020). MadMiner: Machine learning-based inference for particle physics. *Computing and Software for Big Science*, 4:1–25.

Brenier, Y. (1991). Polar factorization and monotone rearrangement of vector-valued functions. *Communications on Pure and Applied Mathematics*, 44:375–417.

Bronshtein, E. M. (1976). ϵ-Entropy of convex sets and functions. *Siberian Mathematical Journal*, 17:393–398.

Bryant, P. E. (2018). *Search for Pair Production of Higgs Bosons in the Four Bottom Quark Final State Using Proton-Proton Collisions at sqrt(s) = 13 TeV with the ATLAS Detector.* PhD thesis, The University of Chicago, Chicago, IL.

Cacciari, M., Salam, G. P., and Soyez, G. (2008). The anti-k_t jet clustering algorithm. *Journal of High Energy Physics*, 2008:063.

Caffarelli, L. A. (1991). Some regularity properties of solutions of Monge Ampère equation. *Communications on Pure and Applied Mathematics*, 44:965–969.

Caffarelli, L. A. (1992a). Boundary regularity of maps with convex potentials. *Communications on Pure and Applied Mathematics*, 45:1141–1151.

Caffarelli, L. A. (1992b). The regularity of mappings with a convex potential. *Journal of the American Mathematical Society*, 5:99–104.

Caffarelli, L. A. (1996). Boundary regularity of maps with convex potentials–II. *The Annals of Mathematics*, 144:453–496.

Caffarelli, L. A. (2000). Monotonicity properties of optimal transportation and the FKG and related inequalities. *Communications in Mathematical Physics*, 214:547–563.

Caffarelli, L. A. and Cabré, X. (1995). *Fully Nonlinear Elliptic Equations.* American Mathematical Soc.

Cai, T., Cheng, J., Craig, N., and Craig, K. (2020). Linearized optimal transport for collider events. *Physical Review D*, 102:116019.

Cairoli, R. (1970). Une inégalité pour martingales à indices multiples et ses applications. *Séminaire de probabilités de Strasbourg*, 4:1–27.

Cairoli, R. and Walsh, J. B. (1975). Stochastic integrals in the plane. *Acta mathematica*, 134:111–183.

Cao, J. and Grigor'yan, A. (2020). Heat kernels and Besov spaces associated with second order divergence form elliptic operators. *Journal of Fourier Analysis and Applications*, 26:3.

Caracciolo, S., Lucibello, C., Parisi, G., and Sicuro, G. (2014). Scaling hypothesis for the Euclidean bipartite matching problem. *Physical Review E*, 90:012118.

Carlier, G., Jimenez, C., and Santambrogio, F. (2008). Optimal transportation with traffic congestion and Wardrop equilibria. *SIAM J. Control Optim.*, 47:1330–1350.

Chaudhuri, K. and Dasgupta, S. (2010). Rates of convergence for the cluster tree. In *Advances in Neural Information Processing Systems 24*, pages 343–351.

Chen, J. (2023). *Statistical Inference under Mixture Models*. ICSA Book Series in Statistics. Springer Nature, Singapore.

Chen, Q. and Fang, Z. (2019). Inference on functionals under first order degeneracy. *Journal of Econometrics*, 210:459–481.

Cheng, K. F. and Chu, C.-K. (2004). Semiparametric density estimation under a two-sample density ratio model. *Bernoulli*, 10:583–604.

Chernozhukov, V., Chetverikov, D., Demirer, M., Duflo, E., Hansen, C., Newey, W., and Robins, J. (2018). Double/debiased machine learning for treatment and structural parameters. *The Econometrics Journal*, 21.

Chernozhukov, V., Galichon, A., Hallin, M., and Henry, M. (2017). Monge–Kantorovich depth, quantiles, ranks and signs. *The Annals of Statistics*, 45:223 – 256.

Chizat, L., Roussillon, P., Léger, F., Vialard, F.-X., and Peyré, G. (2020). Faster Wasserstein Distance Estimation with the Sinkhorn Divergence. *Advances in Neural Information Processing Systems 33*, pages 2257–2269.

Choi, S. and Oh, H. (2021). Improved extrapolation methods of data-driven background estimations in high energy physics. *The European Physical Journal C*, 81:643.

Christmann, A. and Steinwart, I. (2008). *Support Vector Machines*. Springer Science & Business Media.

Christofides, T. C. and Serfling, R. J. (1990a). Maximal inequalities and convergence results for generalized U-statistics. *Journal of Statistical Planning and Inference*, 24:271–286.

Christofides, T. C. and Serfling, R. J. (1990b). Maximal Inequalities for Multidimensionally Indexed Submartingale Arrays. *The Annals of Probability*, 18:630–641.

Cleanthous, G., Georgiadis, A. G., Kerkyacharian, G., Petrushev, P., and Picard, D. (2020). Kernel and wavelet density estimators on manifolds and more general metric spaces. *Bernoulli*, 26:1832 – 1862.

Clozeau, N. and Mattesini, F. (2023). Annealed quantitative estimates for the quadratic 2D-discrete random matching problem. *arXiv preprint arXiv:2303.00353*.

CMS (2008). The CMS experiment at the CERN LHC. *Journal of Instrumentation*, 3:S08004–S08004.

CMS (2012). Observation of a new boson at a mass of 125 GeV with the CMS experiment at the LHC. *Physics Letters B*, 716:30–61.

CMS (2017). Jet energy scale and resolution in the CMS experiment in pp collisions at 8 TeV. *Journal of Instrumentation*, 12:P02014.

CMS (2018). Identification of heavy-flavour jets with the CMS detector in pp collisions at 13 TeV. *Journal of Instrumentation*, 13:P05011.

CMS (2018a). Measurements of Higgs boson properties in the diphoton decay channel in proton-proton collisions at $\sqrt{s} = 13$ TeV. *Journal of High Energy Physics*, 2018:185.

CMS (2018b). Measurements of properties of the Higgs boson in the four-lepton final state at $\sqrt{s} = 13 \text{ TeV}$. *CMS PAS HIG-18-001*.

CMS (2018c). Observation of Higgs boson decay to bottom quarks. *Physical Review Letters*, 121:121801.

CMS (2018d). Observation of the Higgs boson decay to a pair of τ leptons with the CMS detector. *Physics Letters B*, 779:283–316.

CMS (2019). Measurements of properties of the Higgs boson decaying to a W boson pair in pp collisions at s= 13TeV. *Physics letters B*, 791:96–129.

CMS (2022). Search for higgs boson pair production in the four b quark final state in proton-proton collisions at $\sqrt{s}$ = 13 TeV. *arXiv preprint arXiv:2202.09617*.

Cohen, A. (2003). *Numerical Analysis of Wavelet Methods*, volume 32 of *Studies in Mathematics and Its Applications*. North-Holland Publishing Co., Amsterdam.

Cohen, A., Daubechies, I., and Vial, P. (1993). Wavelets on the Interval and Fast Wavelet Transforms. *Applied and Computational Harmonic Analysis*, 1:54–81.

Cohen, D., Kontorovich, A., and Wolfer, G. (2020). Learning discrete distributions with infinite support. *Advances in Neural Information Processing Systems 33*, pages 3942–3951.

Colombo, M. and Fathi, M. (2021). Bounds on optimal transport maps onto log-concave measures. *Journal of Differential Equations*, 271:1007–1022.

Constantine, G. M. and Savits, T. H. (1996). A Multivariate Faà di Bruno Formula with Applications. *Transactions of the American Mathematical Society*, 348:503–520.

Cordero-Erausquin, D. (1999). Sur le transport de mesures périodiques. *Comptes Rendus de l'Académie des Sciences - Series I - Mathematics*, 329:199–202.

Courty, N., Flamary, R., Tuia, D., and Rakotomamonjy, A. (2016). Optimal transport for domain adaptation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 39:1853–1865.

Cover, T. (1968). Estimation by the nearest neighbor rule. *IEEE Transactions on Information Theory*, 14:50–55.

Cranmer, K., Pavez, J., and Louppe, G. (2015). Approximating likelihood ratios with calibrated discriminative classifiers. *arXiv preprint arXiv:1506.02169*.

Csorgo, M. and Revesz, P. (1978). Strong approximations of the quantile process. *The Annals of Statistics*, 6:882–894.

Cuevas, A. (2009). Set estimation: Another bridge between statistics and geometry. *Bol. Estad. Investig. Oper*, 25:71–85.

Cuevas, A. and Fraiman, R. (1997). A plug-in approach to support estimation. *The Annals of Statistics*, 25:2300–2312.

Cuturi, M. (2013). Sinkhorn distances: Lightspeed computation of optimal transport. In *Advances in Neural Information Processing Systems 26*, pages 2292–2300.

Darling, D. A. and Robbins, H. (1967). Confidence sequences for mean, variance, and median. *Proceedings of the National Academy of Sciences*, 58:66–68.

Daubechies, I. (1988). Orthonormal Bases of Compactly Supported Wavelets. *Communications on Pure and Applied Mathematics*, 41:909–996.

Daubechies, I. (1992). *Ten Lectures on Wavelets*. Society for Industrial and Applied Mathematics.

Davies, E. B. (1989). *Heat Kernels and Spectral Theory*. Cambridge University Press.

De Lara, L., González-Sanz, A., and Loubes, J.-M. (2021). A consistent extension of discrete optimal transport maps for machine learning applications. *arXiv preprint arXiv:2102.08644*.

De Philippis, G. and Figalli, A. (2013). Second order stability for the Monge–Ampère equation and strong Sobolev convergence of optimal transport maps. *Analysis & PDE*, 6:993–1000.

De Philippis, G. and Figalli, A. (2014). The Monge–Ampère equation and its link to optimal transportation. *Bulletin of the American Mathematical Society*, 51:527–580.

Deb, N., Bhattacharya, B. B., and Sen, B. (2021). Efficiency lower bounds for distribution-free hotelling-type two-sample tests based on optimal transport. *arXiv preprint arXiv:2104.01986*.

Deb, N., Ghosal, P., and Sen, B. (2021). Rates of Estimation of Optimal Transport Maps using Plug-in Estimators via Barycentric Projections. *Advances in Neural Information Processing Systems 34*, pages 29736–29753.

Deb, N. and Mukherjee, D. (2024). Trade-off Between Dependence and Complexity for Nonparametric Learning – an Empirical Process Approach. *arXiv preprint arXiv:2401.08978*.

Deb, N. and Sen, B. (2021). Multivariate Rank-Based Distribution-Free Nonparametric Testing Using Measure Transportation. *Journal of the American Statistical Association*, 118:197–207.

del Barrio, E., Cuesta-Albertos, J. A., Hallin, M., and Matrán, C. (2020). Center-outward distribution functions, quantiles, ranks, and signs in $\mathbb{R}^d$. *arXiv preprint arXiv:1806.01238*.

del Barrio, E., Cuesta-Albertos, J. A., Matrán, C., and Rodríguez-Rodríguez, J. M. (1999). Tests of goodness of fit based on the L_2-Wasserstein distance. *The Annals of Statistics*, 27:1230–1239.

del Barrio, E., Giné, E., and Utzet, F. (2005). Asymptotics for L^2 functionals of the empirical quantile process, with applications to tests of fit based on weighted Wasserstein distances. *Bernoulli*, 11:131–189.

del Barrio, E., González-Sanz, A., and Loubes, J.-M. (2021). Central limit theorems for general transportation costs. *arXiv preprint arXiv:2102.06379*.

del Barrio, E., González-Sanz, A., and Loubes, J.-M. (2024). Central limit theorems for semi-discrete Wasserstein distances. *Bernoulli*, 30(1):554–580.

del Barrio, E., Gordaliza, P., and Loubes, J.-M. (2019). A central limit theorem for L_p transportation cost with applications to fairness assessment in machine learning. *Information and Inference: A Journal of the IMA*, 8:817–849.

del Barrio, E. and Loubes, J.-M. (2019). Central limit theorems for empirical transportation cost in general dimension. *The Annals of Probability*, 47:926–951.

del Barrio, E., Sanz, A. G., Loubes, J.-M., and Niles-Weed, J. (2023). An improved central limit theorem and fast convergence rates for entropic transportation costs. *SIAM Journal on Mathematics of Data Science*, 5:639–669.

Delalande, A. and Merigot, Q. (2023). Quantitative stability of optimal transport maps under variations of the target measure. *Duke Mathematical Journal*, 172:3321–3357.

Dembinski, H., Kenzie, M., Langenbruch, C., and Schmelling, M. (2022). Custom Orthogonal Weight functions (COWs) for event classification. *Nuclear Instruments and Methods in Physics Research Section A: Accelerators, Spectrometers, Detectors and Associated Equipment*, page 167270.

Denker, M. (1985). *Asymptotic Distribution Theory in Nonparametric Statistics*. Braunschweig-Wiesbaden: Vieweg.

Deshpande, I., Hu, Y.-T., Sun, R., Pyrros, A., Siddiqui, N., Koyejo, S., Zhao, Z., Forsyth, D., and Schwing, A. G. (2019). Max-Sliced Wasserstein distance and its use for GANs. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 10648–10656.

Di Micco, B., Gouzevitch, M., Mazzitelli, J., and Vernieri, C. (2020). Higgs boson potential at colliders: Status and perspectives. *Reviews in Physics*, 5:100045.

Divol, V. (2021). A short proof on the rate of convergence of the empirical measure for the Wasserstein distance. *arXiv preprint arXiv:2101.08126*.

Divol, V., Niles-Weed, J., and Pooladian, A.-A. (2022). Optimal transport map estimation in general function spaces. *arXiv preprint arXiv:2212.03722*.

Do, D., Do, L., McKinley, S. A., Terhorst, J., and Nguyen, X. (2024). Dendrogram of mixing measures: Hierarchical clustering and model selection for finite mixture models. *arXiv preprint arXiv:2403.01684*.

Doob, J. L. (1940). Regularity properties of certain families of chance variables. *Transactions of the American Mathematical Society*, 47:455–486.

Doob, J. L. (1953). *Stochastic Processes*. New York Wiley.

Dudley, R. M. (1969). The speed of mean Glivenko-Cantelli convergence. *The Annals of Mathematical Statistics*, 40:40–50.

Dudley, R. M. (2014). *Uniform Central Limit Theorems*, volume 142. Cambridge University Press.

Dunlop, M. M., Slepčev, D., Stuart, A. M., and Thorpe, M. (2020). Large Data and Zero Noise Limits of Graph-Based Semi-Supervised Learning Algorithms. *Applied and Computational Harmonic Analysis*, 49:655–697.

Durrett, R. (2019). *Probability: Theory and Examples*, volume 49. Cambridge University Press.

Dvoretzky, A., Kiefer, J., and Wolfowitz, J. (1956). Asymptotic minimax character of the sample distribution function and of the classical multinomial estimator. *The Annals of Mathematical Statistics*, 27:642–669.

Efromovich, S. (1999). *Nonparametric Curve Estimation: Methods, Theory, and Applications*. Springer Series in Statistics. Springer-Verlag, New York.

Efron, B. (1971). Forcing a sequential experiment to be balanced. *Biometrika*, 58:403–417.

Efron, B. and Tibshirani, R. J. (1994). *An Introduction to the Bootstrap*. CRC press.

Encinas, L. H. and Masque, J. M. (2003). A short proof of the generalized Faà di Bruno's formula. *Applied Mathematics Letters*, 16:975–979.

Englert, F. and Brout, R. (1964). Broken symmetry and the mass of gauge vector mesons. *Physical Review Letters*, 13:321.

Evans, L. C. (1998). *Partial Differential Equations*, volume 19 of *Graduate Studies in Mathematics*. Springer.

Fan, J. and Hu, T.-C. (1992). Bias correction and higher order kernel functions. *Statistics & Probability Letters*, 13:235–243.

Fatras, K., Zine, Y., Majewski, S., Flamary, R., Gribonval, R., and Courty, N. (2021). Minibatch optimal transport distances; analysis and applications. *arXiv preprint arXiv:2101.01792*.

Fefferman, C. (1971). The multiplier problem for the ball. *The Annals of Mathematics*, 94:330–336.

Figalli, A. (2010). The optimal partial transport problem. *Archive for rational mechanics and analysis*, 195:533–560.

Figalli, A. (2017). *The Monge–Ampère Equation and Its Applications.* European Math. Soc.

Figalli, A. and Glaudo, F. (2021). *An Invitation to Optimal Transport, Wasserstein Distances, and Gradient Flows.* EMS Press.

Finlay, C., Gerolin, A., Oberman, A. M., and Pooladian, A.-A. (2020). Learning normalizing flows from Entropy-Kantorovich potentials. *arXiv preprint arXiv:2006.06033.*

Fix, E. and Hodges, J. (1951). Discriminatory analysis. Nonparametric discrimination; consistency properties. Technical Report 4, Project No. 21-49-004, USAF School of Aviation Medicine, Randolph Field, Texas.

Flamary, R., Courty, N., Gramfort, A., Alaya, M. Z., Boisbunon, A., Chambon, S., Chapel, L., Corenflos, A., Fatras, K., Fournier, N., et al. (2021). POT: Python optimal transport. *Journal of Machine Learning Research*, 22:1–8.

Forrow, A., Hütter, J.-C., Nitzan, M., Rigollet, P., Schiebinger, G., and Weed, J. (2018). Statistical Optimal Transport via Factored Couplings. *The 22nd International Conference on Artificial Intelligence and Statistics*, pages 2454–2465.

Fournier, N. and Guillin, A. (2015). On the rate of convergence in Wasserstein distance of the empirical measure. *Probability Theory and Related Fields*, 162:707–738.

Franke, J. and Runst, T. (1995). Regular Elliptic Boundary Value Problems in Besov-Triebel-Lizorkin Spaces. *Mathematische Nachrichten*, 174:113–149.

Freitag, G., Czado, C., and Munk, A. (2007). A nonparametric test for similarity of marginals—With applications to the assessment of population bioequivalence. *Journal of Statistical Planning and Inference*, 137:697–711.

Freitag, G. and Munk, A. (2005). On Hadamard differentiability in $k-$sample semiparametric models—with applications to the assessment of structural relationships. *Journal of Multivariate Analysis*, 94:123–158.

Freitag, G., Munk, A., and Vogt, M. (2003). Assessing structural relationships between distributions-a quantile process approach based on Mallows distance. In *Recent Advances and Trends in Nonparametric Statistics*, pages 123–137. Elsevier.

Galichon, A. (2016). *Optimal Transport Methods in Economics.* Princeton University Press.

Gangbo, W. and McCann, R. J. (1996). The geometry of optimal transportation. *Acta Mathematica*, 177:113–161.

Garivier, A. (2013). Informational confidence bounds for self-normalized averages and applications. In *2013 IEEE Information Theory Workshop (ITW)*, pages 1–5.

Georgiadis, A. G. and Kyriazis, G. (2023). Duals of Besov and Triebel-Lizorkin spaces associated with operators. *Constructive Approximation*, 57:547–577.

Ghodrati, L. and Panaretos, V. M. (2022). Distribution-on-distribution regression via optimal transport maps. *Biometrika*, 109:957–974.

Ghosal, P. and Sen, B. (2022). Multivariate ranks and quantiles using optimal transport: Consistency, rates and nonparametric testing. *The Annals of Statistics*, 50:1012–1037.

Gigli, N. (2011). On Hölder continuity-in-time of the optimal transport map towards measures along a curve. *Proceedings of the Edinburgh Mathematical Society*, 54:401–409.

Gilbarg, D. and Trudinger, N. S. (2001). *Elliptic Partial Differential Equations of Second Order*. Classics in Mathematics. Springer-Verlag, second edition.

Giné, E. and Guillou, A. (2002). Rates of strong uniform consistency for multivariate kernel density estimators. In *Annales de l'Institut Henri Poincare (B) Probability and Statistics*, volume 38, pages 907–921. Elsevier.

Giné, E. and Koltchinskii, V. (2006). Concentration inequalities and asymptotic results for ratio type empirical processes. *The Annals of Probability*, 34:1143–1216.

Giné, E. and Nickl, R. (2008). A simple adaptive estimator of the integrated square of a density. *Bernoulli*, 14.

Giné, E. and Nickl, R. (2009). Uniform limit theorems for wavelet density estimators. *The Annals of Probability*, 37:1605–1646.

Giné, E. and Nickl, R. (2016). *Mathematical Foundations of Infinite-Dimensional Statistical Models*. Cambridge University Press.

Giordano, M. and Nickl, R. (2020). Consistency of Bayesian inference with Gaussian process priors in an elliptic inverse problem. *Inverse Problems*, 36:085001.

Goldfeld, Z. and Greenewald, K. (2020). Gaussian-smoothed optimal transport: Metric structure and statistical efficiency. In *International Conference on Artificial Intelligence and Statistics*, pages 3327–3337. PMLR.

Goldfeld, Z., Greenewald, K., and Kato, K. (2020). Asymptotic guarantees for generative modeling based on the smooth Wasserstein distance. *Advances in Neural Information Processing Systems 33*, pages 2527–2539.

Goldfeld, Z., Greenewald, K., Niles-Weed, J., and Polyanskiy, Y. (2020). Convergence of smoothed empirical measures with applications to entropy estimation. *IEEE Transactions on Information Theory*, 66:4368–4391.

Goldfeld, Z., Kato, K., Nietert, S., and Rioux, G. (2024a). Limit distribution theory for smooth p-Wasserstein distances. *The Annals of Applied Probability*, 34:2447–2487.

Goldfeld, Z., Kato, K., Rioux, G., and Sadhu, R. (2024b). Limit theorems for entropic optimal transport maps and Sinkhorn divergence. *Electronic Journal of Statistics*, 18:980–1041.

Goldfeld, Z., Kato, K., Rioux, G., and Sadhu, R. (2024c). Statistical inference with regularized optimal transport. *Information and Inference: A Journal of the IMA*, 13:iaad056.

Goldman, M. and Huesmann, M. (2022). A fluctuation result for the displacement in the optimal matching problem. *The Annals of Probability*, 50:1446–1477.

González-Sanz, A. and Hundrieser, S. (2023). Weak limits for empirical entropic optimal transport: Beyond smooth costs. *arXiv preprint arXiv:2305.09745*.

González-Sanz, A., Loubes, J.-M., and Niles-Weed, J. (2022). Weak limits of entropy regularized Optimal Transport; potentials, plans and divergences. *arXiv preprint arXiv:2207.07427*.

Gozlan, N. and Léonard, C. (2010). Transport inequalities. A survey. *Markov Processes and Related Fields*, 16:635–736.

Grafakos, L. (2008). *Classical Fourier Analysis*. Graduate Texts in Mathematics. Springer-Verlag, second edition.

Grafakos, L. (2009). *Modern Fourier Analysis*. Graduate Texts in Mathematics. Springer-Verlag.

Greengard, P., Hoskins, J. G., Marshall, N. F., and Singer, A. (2022). On a linearization of quadratic Wasserstein distance. *arXiv preprint arXiv:2201.13386*.

Gretton, A., Borgwardt, K. M., Rasch, M. J., Schölkopf, B., and Smola, A. (2012). A kernel two-sample test. *Journal of Machine Learning Research*, 13:723–773.

Groppe, M. and Hundrieser, S. (2023). Lower Complexity Adaptation for Empirical Entropic Optimal Transport. *arXiv preprint arXiv:2306.13580*.

Grüter, M. and Widman, K.-O. (1982). The Green function for uniformly elliptic equations. *Manuscripta Mathematica*, 37:303–342.

Guittet, K. (2003). On the time-continuous mass transport problem and its approximation by augmented Lagrangian techniques. *SIAM Journal on Numerical Analysis*, 41:382–399.

Gunsilius, F. (2022). On the convergence rate of potentials of Brenier maps. *Econometric Theory*, 38:381–417.

Gunsilius, F. and Xu, Y. (2021). Matching for causal effects via multimarginal unbalanced optimal transport. *arXiv preprint arXiv:2112.04398*.

Guntuboyina, A. and Sen, B. (2012). L_1 covering numbers for uniformly bounded convex functions. In *Proceedings of the 25th Annual Conference on Learning Theory*, volume 23, pages 12.1–12.13. JMLR Workshop and Conference Proceedings.

Guo, F. R. and Richardson, T. S. (2020). Chernoff-type Concentration of Empirical Probabilities in Relative Entropy. *IEEE Transactions on Information Theory*, 67:549–558.

Guo, H. and Kou, J. (2019). Strong Uniform Convergence Rates of Wavelet Density Estimators with Size-Biased Data. *Journal of Function Spaces*, 2019.

Györfi, L., Kohler, M., Krzyzak, A., and Walk, H. (2006). *A Distribution-Free Theory of Nonparametric Regression*. Springer Science & Business Media.

Hagedorn, R. (1964). *Relativistic Kinematics: A Guide to the Kinematic Problems of High-Energy Physics*. W.A. Benjamin.

Hallin, M. (2022). Measure transportation and statistical decision theory. *Annual Review of Statistics and Its Application*, 9:401–424.

Hallin, M., del Barrio, E., Cuesta-Albertos, J., and Matrán, C. (2021). Distribution and quantile functions, ranks and signs in dimension d: A measure transportation approach. *The Annals of Statistics*, 49:1139 – 1165.

Hallin, M., Mordant, G., and Segers, J. (2021). Multivariate goodness-of-fit tests based on Wasserstein distance. *Electronic Journal of Statistics*, 15:1328–1371.

Han, Y., Jiao, J., and Weissman, T. (2015). Minimax estimation of discrete distributions under ℓ_1 loss. *IEEE Transactions on Information Theory*, 61:6343–6354.

Han, Y., Jiao, J., Weissman, T., and Wu, Y. (2020). Optimal rates of entropy estimation over Lipschitz balls. *The Annals of Statistics*, 48:3228–3250.

Hanley, J. A. and McNeil, B. J. (1982). The meaning and use of the area under a receiver operating characteristic (ROC) curve. *Radiology*, 143:29–36.

Harchaoui, Z., Liu, L., and Pal, S. (2020). Asymptotics of discrete Schrödinger bridges via chaos decomposition. *arXiv preprint arXiv:2011.08963*.

Härdle, W., Kerkyacharian, G., Picard, D., and Tsybakov, A. (2012). *Wavelets, Approximation, and Statistical Applications*, volume 129. Springer Science & Business Media.

Heinrich, J. and Lyons, L. (2007). Systematic errors. *Annu. Rev. Nucl. Part. Sci.*, 57:145–169.

Hendriks, H. (1990). Nonparametric estimation of a probability density on a Riemannian manifold using Fourier expansions. *The Annals of Statistics*, 18:832–849.

Higgs, P. W. (1964). Broken symmetries and the masses of gauge bosons. *Physical Review Letters*, 13:508.

Hiriart-Urruty, J.-B. and Lemaréchal, C. (2004). *Fundamentals of Convex Analysis*. Springer Science & Business Media.

Ho, J., Jain, A., and Abbeel, P. (2020). Denoising Diffusion Probabilistic Models. In *Advances in Neural Information Processing Systems*, volume 33, pages 6840–6851.

Ho, N., Huynh, V., Phung, D., and Jordan, M. (2019). Probabilistic multilevel clustering via composite transportation distance. In *The 22nd International Conference on Artificial Intelligence and Statistics*, pages 3149–3157. PMLR.

Ho, N., Nguyen, X., Yurochkin, M., Bui, H. H., Huynh, V., and Phung, D. (2017). Multilevel clustering via Wasserstein means. In *International Conference on Machine Learning*, pages 1501–1509. PMLR.

Ho, N., Yang, C.-Y., and Jordan, M. I. (2022). Convergence rates for Gaussian mixtures of experts. *Journal of Machine Learning Research*, 23:1–81.

Hörmander, L. (2007). *The Analysis of Linear Partial Differential Operators III: Pseudo-Differential Operators*. Classics in Mathematics, The Analysis of Linear Partial Differential Operators. Springer-Verlag.

Howard, S. R. and Ramdas, A. (2022). Sequential estimation of quantiles with applications to a/b testing and best-arm identification. *Bernoulli*, 28:1704–1728.

Howard, S. R., Ramdas, A., McAuliffe, J., and Sekhon, J. (2020). Time-uniform Chernoff bounds via nonnegative supermartingales. *Probability Surveys*, 17:257–317.

Howard, S. R., Ramdas, A., McAuliffe, J., and Sekhon, J. (2021). Uniform, nonparametric, non-asymptotic confidence sequences. *The Annals of Statistics*, 49:1055–1080.

Huang, Z. and Sen, B. (2023). Multivariate symmetry: Distribution-free testing via optimal transport. *arXiv preprint arXiv:2305.01839*.

Hundrieser, S., Klatt, M., Staudt, T., and Munk, A. (2022). A unifying approach to distributional limits for empirical optimal transport. *Bernoulli (To appear)*.

Hundrieser, S., Staudt, T., and Munk, A. (2022). Empirical Optimal Transport between Different Measures Adapts to Lower Complexity. *Annales de l'Institut Henri Poincaré (To appear)*.

Hütter, J.-C. and Rigollet, P. (2021). Minimax rates of estimation for smooth optimal transport maps. *The Annals of Statistics*, 49:1166–1194.

Huynh, V., Ho, N., Dam, N., Nguyen, X., Yurochkin, M., Bui, H., and Phung, D. (2021). On efficient multilevel clustering via Wasserstein distances. *Journal of Machine Learning Research*, 22:1–43.

Imaizumi, M., Ota, H., and Hamaguchi, T. (2022). Hypothesis test and confidence analysis with wasserstein distance on general dimension. *Neural Computation*, 34:1448–1487.

Ivanoff, B. G. and Merzbach, E. (1999). *Set-Indexed Martingales*, volume 85. CRC Press.

Jacobs, M. and Léger, F. (2020). A fast approach to optimal transport: The back-and-forth method. *Numerische Mathematik*, 146:513–544.

Jamieson, K., Malloy, M., Nowak, R., and Bubeck, S. (2014). Lil'UCB : An Optimal Exploration Algorithm for Multi-Armed Bandits. In *Conference on Learning Theory*, pages 423–439.

Jordan, R., Kinderlehrer, D., and Otto, F. (1998). The variational formulation of the Fokker–Planck equation. *SIAM Journal on Mathematical Analysis*, 29:1–17.

Josien, M. (2019). Decomposition and pointwise estimates of periodic Green functions of some elliptic equations with periodic oscillatory coefficients. *Asymptotic Analysis*, 112:227–246.

Kallenberg, O. (2006). *Probabilistic Symmetries and Invariance Principles*. Springer Science & Business Media.

Kamath, S., Orlitsky, A., Pichapati, D., and Suresh, A. T. (2015). On Learning Distributions from their Samples. In *Proceedings of The 28th Conference on Learning Theory*, pages 1066–1100. PMLR.

Kantorovich, L. V. (1942). On the translocation of masses. In *Dokl. Akad. Nauk. USSR (NS)*, volume 37, pages 199–201.

Kantorovich, L. V. (1948). On a problem of Monge. In *CR (Doklady) Acad. Sci. URSS (NS)*, volume 3, pages 225–226.

Karampatziakis, N., Mineiro, P., and Ramdas, A. (2021). Off-policy Confidence Sequences. In *Proceedings of the 38th International Conference on Machine Learning*, pages 5301–5310. PMLR 139.

Kasieczka, G., Nachman, B., Schwartz, M. D., and Shih, D. (2021). Automating the ABCD method with machine learning. *Physical Review D*, 103:035021.

Kaufmann, E., Cappé, O., and Garivier, A. (2016). On the complexity of best-arm identification in multi-armed bandit models. *Journal of Machine Learning Research*, 17:1–42.

Kaufmann, E. and Koolen, W. M. (2021). Mixture martingales revisited with applications to sequential tests and confidence intervals. *Journal of Machine Learning Research*, 22:246–1.

Kennedy, E. H. (2022). Semiparametric doubly robust targeted double machine learning: A review. *arXiv preprint arXiv:2203.06469*.

Kennedy, E. H., Balakrishnan, S., and G'Sell, M. (2020). Sharp instruments for classifying compliers and generalizing causal effects. *The Annals of Statistics*, 48:2008–2030.

Kerkyacharian, G. and Petrushev, P. (2015). Heat kernel based decomposition of spaces of distributions in the framework of Dirichlet spaces. *Transactions of the American Mathematical Society*, 367:121–189.

Kerkyacharian, G. and Picard, D. (1992). Density estimation in Besov spaces. *Statistics & Probability Letters*, 13:15–24.

Kerkyacharian, G. and Picard, D. (1996). Estimating nonquadratic functionals of a density using Haar wavelets. *The Annals of Statistics*, 24:485–507.

Kim, I., Balakrishnan, S., and Wasserman, L. (2020). Robust multivariate nonparametric tests via projection averaging. *The Annals of Statistics*, 48:3417–3441.

Klatt, M., Munk, A., and Zemel, Y. (2022). Limit laws for empirical optimal solutions in random linear programs. *Annals of Operations Research*, 315:251–278.

Klatt, M., Tameling, C., and Munk, A. (2020). Empirical regularized optimal transport: Statistical theory and applications. *SIAM Journal on Mathematics of Data Science*, 2:419–443.

Klein, N., Orellana, J., Brincat, S. L., Miller, E. K., and Kass, R. E. (2020). Torus graphs for multivariate phase coupling analysis. *The Annals of Applied Statistics*, 14:635–660.

Knott, M. and Smith, C. S. (1984). On the optimal mapping of distributions. *Journal of Optimization Theory and Applications*, 43:39–49.

Kobyzev, I., Prince, S. J., and Brubaker, M. A. (2020). Normalizing flows: An introduction and review of current methods. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 43:3964–3979.

Kolouri, S., Nadjahi, K., Simsekli, U., Badeau, R., and Rohde, G. (2019). Generalized sliced Wasserstein distances. In *Advances in Neural Information Processing Systems 32*, pages 261–272.

Kolouri, S., Park, S. R., Thorpe, M., Slepcev, D., and Rohde, G. K. (2017). Optimal mass transport: Signal processing and machine-learning applications. *IEEE Signal Processing Magazine*, 34:43–59.

Komiske, Metodiev, E., and Thaler, J. (2022). EnergyFlow Python Package. https://energyflow.network/.

Komiske, P. T., Mastandrea, R., Metodiev, E. M., Naik, P., and Thaler, J. (2020). Exploring the space of jets with CMS open data. *Physical Review D*, 101:034009.

Komiske, P. T., Metodiev, E. M., and Thaler, J. (2019). Metric space of collider events. *Physical Review Letters*, 123:041801.

Komiske, P. T., Metodiev, E. M., and Thaler, J. (2020). The hidden geometry of particle collisions. *Journal of High Energy Physics*, 2020:6.

Kpotufe, S. (2017). Lipschitz density-ratios, structured data, and data-driven tuning. In *Artificial Intelligence and Statistics*, pages 1320–1328. PMLR.

Krause, C. and Shih, D. (2023a). Accelerating accurate simulations of calorimeter showers with normalizing flows and probability density distillation. *Physical Review D*, 107:113004.

Krause, C. and Shih, D. (2023b). Fast and accurate simulations of calorimeter showers with normalizing flows. *Physical Review D*, 107:113003.

Krishnamurthy, A., Kandasamy, K., Poczos, B., and Wasserman, L. (2014). Nonparametric estimation of Renyi divergence and friends. In *International Conference on Machine Learning*, pages 919–927.

Krishnamurthy, A., Kandasamy, K., Poczos, B., and Wasserman, L. (2015). On estimating L_2^2 divergence. In *Proceedings of the Eighteenth International Conference on Artificial Intelligence and Statistics*, pages 498–506.

Kuchibhotla, A. K. and Chakrabortty, A. (2022). Moving beyond sub-Gaussianity in high-dimensional statistics: Applications in covariance estimation and linear regression. *Information and Inference: A Journal of the IMA*, 11:1389–1456.

Kuhn, H. W. (1955). The Hungarian method for the assignment problem. *Naval Research Logistics Quarterly*, 2:83–97.

Lai, T. L. (1976). On confidence sequences. *The Annals of Statistics*, 4:265–280.

Laurent, B. (1996). Efficient estimation of integral functionals of a density. *The Annals of Statistics*, 24:659–681.

Le, T., Yamada, M., Fukumizu, K., and Cuturi, M. (2019). Tree-sliced variants of Wasserstein distances. In *Advances in Neural Information Processing Systems 32*, pages 12283–12294.

Ledoux, M. (2019). On optimal matching of gaussian samples. *Journal of Mathematical Sciences*, 238:495–522.

Lee, A. J. (1990). *U-Statistics: Theory and Practice.* CRC Press.

Lee, J., Dabagia, M., Dyer, E., and Rozell, C. (2019). Hierarchical optimal transport for multimodal distribution alignment. In *Advances in Neural Information Processing Systems*, pages 13453–13463.

Lei, J. (2020). Convergence and concentration of empirical measures under Wasserstein distance in unbounded functional spaces. *Bernoulli*, 26:767–798.

Lei, J. (2021). Network representation using graph root distributions. *The Annals of Statistics*, 49:745–768.

Leoni, G. (2017). *A First Course in Sobolev Spaces.* American Mathematical Soc.

Levy, B. and Schwindt, E. (2018). Notions of optimal transport theory and how to implement them on a computer. *Computers & Graphics*, 72:135–148.

Lhéritier, A. and Cazals, F. (2018). A sequential non-parametric multivariate two-sample test. *IEEE Transactions on Information Theory*, 64:3361–3370.

Li, Q.-R., Santambrogio, F., and Wang, X.-J. (2014). Regularity in Monge's mass transfer problem. *Journal de Mathématiques Pures et Appliquées*, 102:1015–1040.

Liang, T. (2019). On the minimax optimality of estimating the Wasserstein metric. *arXiv preprint arXiv:1908.10324*.

Liang, T. (2021). How well generative adversarial networks learn distributions. *Journal of Machine Learning Research*, 22:1–41.

Liero, M., Mielke, A., and Savaré, G. (2018). Optimal entropy-transport problems and a new Hellinger–Kantorovich distance between positive measures. *Inventiones mathematicae*, 211:969–1117.

Lin, T., Cuturi, M., and Jordan, M. (2024). A specialized semismooth Newton method for kernel-based optimal transport.

Littman, W., Stampacchia, G., and Weinberger, H. F. (1963). Regular points for elliptic equations with discontinuous coefficients. *Annali della Scuola Normale Superiore di Pisa-Classe di Scienze*, 17:43–77.

Loeper, G. (2006). Uniqueness of the solution to the Vlasov–Poisson system with bounded density. *Journal de Mathématiques Pures et Appliquées*, 86:68–79.

Loeper, G. and Rapetti, F. (2005). Numerical solution of the Monge–Ampère equation by a Newton's algorithm. *Comptes rendus. Mathématique*, 340:319–324.

Löfström, J. (1992). Interpolation of boundary value problems of Neumann type on smooth domains. *Journal of the London Mathematical Society*, 2:499–516.

Lunardi, A. (2018). *Interpolation Theory*. Edizioni della Normale Pisa.

Lyons, L. (1986). *Statistics for Nuclear and Particle Physicists*. Cambridge University Press.

Ma, X.-N., Trudinger, N. S., and Wang, X.-J. (2005). Regularity of Potential Functions of the Optimal Transportation Problem. *Archive for Rational Mechanics and Analysis*, 177:151–183.

Maillard, O.-A. (2021). Local Dvoretzky–Kiefer–Wolfowitz Confidence Bands. *Mathematical Methods of Statistics*, 30:16–46.

Makkuva, A., Taghvaei, A., Oh, S., and Lee, J. (2020). Optimal transport mapping via input convex neural networks. In *International Conference on Machine Learning*, pages 6672–6681. PMLR.

Manole, T., Balakrishnan, S., Niles-Weed, J., and Wasserman, L. (2021). Plugin estimation of smooth optimal transport maps. *The Annals of Statistics (To appear)*.

Manole, T., Bryant, P., Alison, J., Kuusela, M., and Wasserman, L. (2022). Background modeling for double Higgs boson production: Density ratios and optimal transport. *arXiv preprint arXiv:2208.02807*.

Masry, E. (1997). Multivariate probability density estimation by wavelet methods: Strong consistency and rates for stationary time series. *Stochastic processes and their applications*, 67:177–193.

Massart, P. (1990). The tight constant in the Dvoretzky-Kiefer-Wolfowitz inequality. *The Annals of Probability*, pages 1269–1283.

Mazumder, R., Choudhury, A., Iyengar, G., and Sen, B. (2019). A computational framework for multivariate convex regression and its variants. *Journal of the American Statistical Association*, 114:318–331.

Mena, G. and Weed, J. (2019). Statistical bounds for entropic optimal transport: Sample complexity and the central limit theorem. In *Advances in Neural Information Processing Systems 32*.

Mérigot, Q. (2011). A multiscale approach to optimal transport. In *Computer Graphics Forum*, volume 30, pages 1583–1592.

Mérigot, Q., Delalande, A., and Chazal, F. (2020). Quantitative stability of optimal transport maps and linearization of the 2-Wasserstein space. In *International Conference on Artificial Intelligence and Statistics*, pages 3186–3196. PMLR.

Merzbach, E. (2003). An introduction to the general theory of set-indexed martingales. In *Topics in Spatial Stochastic Processes*, pages 41–84. Springer.

Meyer, Y. (1991). Ondelettes sur l'intervalle. *Revista Matematica Iberoamericana*, 7:115–133.

Miranda, C. (2013). *Partial Differential Equations of Elliptic Type.* Springer-Verlag.

Mohri, M., Rostamizadeh, A., and Talwalkar, A. (2018). *Foundations of Machine Learning.* MIT Press.

Monge, G. (1781). Mémoire sur la théorie des déblais et des remblais. *Histoire de l'Académie Royale des Sciences de Paris.*

Müller, A. (1997). Integral probability metrics and their generating classes of functions. *Advances in Applied Probability*, 29:429–443.

Munk, A. and Czado, C. (1998). Nonparametric validation of similar distributions and assessment of goodness of fit. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 60:223–241.

Nadjahi, K., De Bortoli, V., Durmus, A., Badeau, R., and Şimşekli, U. (2020). Approximate Bayesian computation with the Sliced-Wasserstein distance. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 5470–5474.

Nath, J. S. and Jawanpuria, P. (2020). Statistical optimal transport posed as learning kernel mean embedding. In *Advances in Neural Information Processing Systems 34*, pages 17334–17345.

Nguyen, K., Ho, N., Pham, T., and Bui, H. (2020). Distributional Sliced-Wasserstein and applications to Generative Modeling. In *International Conference on Learning Representations*.

Nguyen, K., Nguyen, D., Pham, T., and Ho, N. (2022). Improving mini-batch optimal transport via partial transportation. In *Proceedings of the 38th International Conference on Machine Learning*, pages 16656–16690. PMLR 139.

Nguyen, X. (2013). Convergence of latent mixing measures in finite and infinite mixture models. *The Annals of Statistics*, 41:370–400.

Nguyen, X., Wainwright, M. J., and Jordan, M. I. (2010). Estimating divergence functionals and the likelihood ratio by convex risk minimization. *IEEE Transactions on Information Theory*, 56:5847–5861.

Nickl, R., Van De Geer, S., and Wang, S. (2020). Convergence rates for penalized least squares estimators in PDE constrained regression problems. *SIAM/ASA Journal on Uncertainty Quantification*, 8:374–413.

Niles-Weed, J. and Berthet, Q. (2022). Minimax estimation of smooth densities in Wasserstein distance. *The Annals of Statistics*, 50:1519–1540.

Niles-Weed, J. and Rigollet, P. (2022). Estimation of Wasserstein distances in the spiked transport model. *Bernoulli*, 28:2663–2688.

Nishiyama, Y. (2011). Impossibility of weak convergence of kernel density estimators to a non-degenerate law in $L^2(\mathbb{R}^d)$. *Journal of Nonparametric Statistics*, 23:129–135.

Onken, D., Fung, S. W., Li, X., and Ruthotto, L. (2021). OR-Flow: Fast and accurate continuous normalizing flows via optimal transport. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pages 9223–9232.

Panaretos, V. M. and Zemel, Y. (2019a). *An Invitation to Statistics in Wasserstein Space*. Springer Nature.

Panaretos, V. M. and Zemel, Y. (2019b). Statistical aspects of Wasserstein distances. *Annual Review of Statistics and Its Application*, 6:405–431.

Paty, F.-P. and Cuturi, M. (2019). Subspace Robust Wasserstein Distances. In *Proceedings of the 36th International Conference on Machine Learning*, pages 5072–5081.

Pele, O. and Werman, M. (2008). A linear time histogram metric for improved sift matching. In *European Conference on Computer Vision*, pages 495–508. Springer.

Peleg, S., Werman, M., and Rom, H. (1989). A unified approach to the change of resolution: Space and gray-level. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 11:739–742.

Perrot, M., Courty, N., Flamary, R., and Habrard, A. (2016). Mapping Estimation for Discrete Optimal Transport. In Lee, D. D., Sugiyama, M., Luxburg, U. V., Guyon, I., and Garnett, R., editors, *Advances in Neural Information Processing Systems 29*, pages 4197–4205.

Petersen, A. and Müller, H.-G. (2019). Fréchet regression for random objects with Euclidean predictors. *The Annals of Statistics*, 47:691–719.

Peyré, G. and Cuturi, M. (2019). Computational optimal transport. *Foundations and Trends® in Machine Learning*, 11:355–607.

Peyre, R. (2018). Comparison between W_2 distance and $\dot{H}^{-1}$ norm, and localisation of Wasserstein distance. *ESAIM: Control, Optimisation and Calculus of Variations*, 24:1489–1501.

Pivk, M. and Le Diberder, F. (2005). sPlot: A statistical tool to unfold data distributions. *Nuclear Instruments and Methods in Physics Research Section A: Accelerators, Spectrometers, Detectors and Associated Equipment*, 555:356–369.

Placakyte, R. (2011). Parton Distribution Functions. *arXiv preprint arXiv:1111.5452*.

Póczos, B. and Schneider, J. (2012). Nonparametric estimation of conditional information and divergences. In *Proceedings of the Fifteenth International Conference on Artificial Intelligence and Statistics*, volume 22, pages 914–923. PMLR.

Pollard, C. and Windischhofer, P. (2022). Transport away your problems: Calibrating stochastic simulations with optimal transport. *Nuclear Instruments and Methods in Physics Research Section A: Accelerators, Spectrometers, Detectors and Associated Equipment*, 1027:166119.

Pollard, D. (1981). Limit theorems for empirical processes. *Zeitschrift für Wahrscheinlichkeitstheorie und Verwandte Gebiete*, 57:181–195.

Pollard, D. (2002). *A User's Guide to Measure Theoretic Probability*. Cambridge University Press.

Polyanskiy, Y. and Wu, Y. (2016). Wasserstein Continuity of Entropy and Outer Bounds for Interference Channels. *IEEE Transactions on Information Theory*, 62:3992–4002.

Ponnoprat, D., Okano, R., and Imaizumi, M. (2024). Uniform confidence band for optimal transport map on one-dimensional data. *Electronic Journal of Statistics*, 18:515–552.

Pooladian, A.-A., Divol, V., and Niles-Weed, J. (2023). Minimax estimation of discontinuous optimal transport maps: The semi-discrete case. *arXiv preprint arXiv:2301.11302*.

Pooladian, A.-A. and Niles-Weed, J. (2021). Entropic estimation of optimal transport maps. *arXiv preprint arXiv:2109.12004*.

Qin, J. (1998). Inferences for case-control and semiparametric two-sample density ratio models. *Biometrika*, 85:619–630.

Rabin, J., Peyré, G., Delon, J., and Bernot, M. (2011). Wasserstein barycenter and its application to texture mixing. In *International Conference on Scale Space and Variational Methods in Computer Vision*, pages 435–446. Springer.

Rakotomamonjy, A., Flamary, R., Gasso, G., Alaya, M. Z., Berar, M., and Courty, N. (2022). Optimal Transport for Conditional Domain Matching and Label Shift. *Machine Learning*, 111:1651–1670.

Ramdas, A. and Manole, T. (2023). Randomized and exchangeable improvements of Markov's, Chebyshev's and Chernoff's inequalities. *arXiv preprint arXiv:2304.02611*.

Ramdas, A., Ruf, J., Larsson, M., and Koolen, W. (2020). Admissible anytime-valid sequential inference must rely on nonnegative martingales. *arXiv preprint arXiv:2009.03167*.

Ramdas, A., Trillos, N., and Cuturi, M. (2017). On Wasserstein two-sample testing and related families of nonparametric tests. *Entropy*, 19:47.

Read, A. L. (1999). Linear interpolation of histograms. *Nuclear Instruments and Methods in Physics Research Section A: Accelerators, Spectrometers, Detectors and Associated Equipment*, 425:357–360.

Redko, I., Courty, N., Flamary, R., and Tuia, D. (2019). Optimal transport for multi-source domain adaptation under target shift. In *The 22nd International Conference on Artificial Intelligence and Statistics*, pages 849–858. PMLR.

Reiss, R.-D. (2012). *A course on point processes*. Springer Science & Business Media.

Revuz, D. and Yor, M. (2013). *Continuous Martingales and Brownian Motion*. Springer Science & Business Media.

Rigollet, P. and Hütter, J.-C. (2015). High dimensional statistics. *Lecture notes*.

Rippl, T., Munk, A., and Sturm, A. (2016). Limit laws of the empirical Wasserstein distance: Gaussian distributions. *Journal of Multivariate Analysis*, 151:90–109.

Robbins, H. and Siegmund, D. (1969). Probability distributions related to the law of the iterated logarithm. *Proceedings of the National Academy of Sciences*, 62:11–13.

Robins, J., Li, L., Tchetgen, E., and van der Vaart, A. W. (2009). Quadratic semiparametric von Mises calculus. *Metrika*, 69:227–247.

Rosenthal, H. (1972). On the span in L^p of sequences of independent random variables (II). In *Proceedings of the Sixth Berkeley Symposium on Mathematical Statistics and Probability: Held at the Statistical Laboratory, University of California, June 21-July 18, 1970. Probability Theory*, pages 149–167. Univ of California Press.

Rosenthal, H. P. (1970). On the subspaces of L^p $(p > 2)$ spanned by sequences of independent random variables. *Israel Journal of Mathematics*, 8:273–303.

Rubenstein, P. K., Bousquet, O., Djolonga, J., Riquelme, C., and Tolstikhin, I. (2019). Practical and consistent estimation of f-divergences. *Advances in Neural Information Processing Systems 32*, page 4070–4080.

Rzeszut, M. and Trojan, B. (2020). Concrete representation of atomic F_4 filtrations. *arXiv preprint arXiv:2001.00196*.

Sadhu, R., Goldfeld, Z., and Kato, K. (2021). Limit distribution theory for the smooth 1-wasserstein distance with applications. *arXiv preprint arXiv:2107.13494*.

Sadhu, R., Goldfeld, Z., and Kato, K. (2023). Stability and statistical inference for semidiscrete optimal transport maps. *arXiv preprint arXiv:2303.10155*.

Sakuma, T. and McCauley, T. (2014). Detector and event visualization with SketchUp at the CMS experiment. In *Journal of Physics: Conference Series*, volume 513, page 022032. IOP Publishing.

Saloff-Coste, L. (2010). The heat kernel and its estimates. *Probabilistic Approach to Geometry*, 57:405–436.

Samworth, R. and Johnson, O. (2005). The empirical process in Mallows distance, with application to goodness-of-fit tests. *arXiv preprint arXiv:math/0504424*.

Santambrogio, F. (2015). *Optimal Transport for Applied Mathematicians: Calculus of Variations, PDEs, and Modeling*, volume 87. Birkhäuser.

Schmeisser, H.-J. and Triebel, H. (1987). *Topics in Fourier Analysis and Function Spaces*. Wiley.

Schölkopf, B. and Smola, A. J. (2018). *Learning with Kernels: Support Vector Machines, Regularization, Optimization, and Beyond*. Adaptive Computation and Machine Learning Series. MIT Press.

Seeley, R. (1972). Interpolation in L^p with boundary conditions. *Studia Mathematica*, 44:47–60.

Segers, J. (2022). Graphical and uniform consistency of estimated optimal transport plans. *arXiv preprint arXiv:2208.02508*.

Seijo, E. and Sen, B. (2011). Nonparametric least squares estimation of a multivariate convex regression function. *The Annals of Statistics*, 39:1633–1657.

Serfling, R. J. (2009). *Approximation Theorems of Mathematical Statistics*, volume 162. John Wiley & Sons.

Shao, J. and Tu, D. (2012). *The Jackknife and Bootstrap*. Springer Science & Business Media.

Shi, H., Drton, M., and Han, F. (2020). Distribution-free consistent independence tests via center-outward ranks and signs. *Journal of the American Statistical Association*, 0:1–16.

Shin, J., Ramdas, A., and Rinaldo, A. (2021). On the bias, risk and consistency of sample means in multi-armed bandits. *SIAM Journal on Mathematics of Data Science*, 3:1278–1300.

Shorack, G. R. and Wellner, J. A. (2009). *Empirical Processes with Applications to Statistics*. Society for Industrial and Applied Mathematics.

Silverman, B. W. (1978). Weak and strong uniform consistency of the kernel estimate of a density and its derivatives. *The Annals of Statistics*, 6:177–184.

Silverman, B. W. and Jones, M. C. (1989). E. Fix and J.L. Hodges (1951): An important contribution to nonparametric discriminant analysis and density estimation: Commentary on fix and hodges (1951). *International Statistical Review*, pages 233–238.

Singh, S. and Póczos, B. (2014). Generalized exponential concentration inequality for Rényi divergence estimation. In *Proceedings of the Thirty-First International Conference on Machine Learning*, pages 333–341. PMLR.

Singh, S. and Póczos, B. (2019). Minimax distribution estimation in Wasserstein distance. *arXiv preprint arXiv:1802.08855*.

Singh, S., Uppal, A., Li, B., Li, C.-L., Zaheer, M., and Poczos, B. (2018). Nonparametric Density Estimation under Adversarial Losses. In *Advances in Neural Information Processing Systems 31*.

Sisson, S. A., Fan, Y., and Beaumont, M. (2018). *Handbook of Approximate Bayesian Computation*. Chapman and Hall/CRC.

Smirnov, N. V. (1944). Approximate laws of distribution of random variables from empirical data. *Uspekhi Matematicheskikh Nauk*, 10:179–206.

Sommerfeld, M. and Munk, A. (2018). Inference for empirical Wasserstein distances on finite spaces. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 80:219–238.

Sommerfeld, M., Schrieber, J., Zemel, Y., and Munk, A. (2019). Optimal Transport: Fast Probabilistic Approximation with Exact Solvers. *Journal of Machine Learning Research*, 20:1–23.

Sricharan, K., Raich, R., and Hero III, A. O. (2010). Empirical estimation of entropy functionals with confidence. *arXiv preprint arXiv:1012.4188*.

Sriperumbudur, B. K., Fukumizu, K., Gretton, A., Schölkopf, B., and Lanckriet, G. R. G. (2012). On the empirical estimation of integral probability metrics. *Electronic Journal of Statistics*, 6:1550–1599.

Stampacchia, G. (1965). Le problème de Dirichlet pour les équations elliptiques du second ordre à coefficients discontinus. In *Annales de l'Institut Fourier*, volume 15, pages 189–257.

Staudt, T. and Hundrieser, S. (2023). Convergence of Empirical Optimal Transport in Unbounded Settings. *arXiv preprint arXiv:2306.11499*.

Staudt, T., Hundrieser, S., and Munk, A. (2022). On the Uniqueness of Kantorovich Potentials. *arXiv preprint arXiv:2201.08316*.

Stein, E. M. and Shakarchi, R. (2009). *Real Analysis: Measure Theory, Integration, and Hilbert Spaces*. Princeton University Press.

Stein, E. M. and Weiss, G. (1971). *Introduction to Fourier Analysis on Euclidean Spaces*. Princeton University Press.

Steinerberger, S. (2016). Directional Poincaré inequalities along mixing flows. *Arkiv för Matematik*, 54:555–569.

Stout, W. F. (1970). The Hartman-Wintner law of the iterated logarithm for martingales. *The Annals of Mathematical Statistics*, 41:2158–2160.

Strobl, F. (1995). On the reversed sub-martingale property of empirical discrepancies in arbitrary sample spaces. *Journal of Theoretical Probability*, 8:825–831.

Stromme, A. J. (2023). Minimum intrinsic dimension scaling for entropic optimal transport. *arXiv preprint arXiv:2306.03398*.

Stupfler, G. (2014). On the weak convergence of kernel density estimators in L^p spaces. *Journal of Nonparametric Statistics*, 26:721–735.

Stupfler, G. (2016). On the weak convergence of the kernel density estimator in the uniform topology. *Electronic Communications in Probability*, 21:1 – 13.

Sturm (1996). Analysis on local Dirichlet spaces-III. Poincaré and parabolic Harnack inequality. *J. Math. Pures Appl.*, pages 273–297.

Taira, K. (2016). *Analytic Semigroups and Semilinear Initial Boundary Value Problems*. Cambridge University Press.

Talagrand, M. (1992). The Ajtai-Komlos-Tusnady matching theorem for general measures. In Dudley, R. M., Hahn, M. G., and Kuelbs, J., editors, *Probability in Banach Spaces, 8: Proceedings of the Eighth International Conference*, pages 39–54. Birkhäuser Boston, Boston, MA.

Tameling, C., Sommerfeld, M., and Munk, A. (2019). Empirical optimal transport on countable metric spaces: Distributional limits and statistical applications. *The Annals of Applied Probability*, 29:2744–2781.

Tolstikhin, I. O., Sriperumbudur, B. K., and Schölkopf, B. (2016). Minimax estimation of maximum mean discrepancy with radial kernels. In *Advances in Neural Information Processing Systems 30*, pages 1930–1938.

Triebel, H. (1983). *Theory of Function Spaces*. Modern Birkhäuser Classics. Birkhäuser Basel.

Triebel, H. (1995). *Interpolation Theory, Function Spaces*. Johann Ambrosius Barth.

Triebel, H. (2006). *Theory of Function Spaces III*. Monographs in Mathematics, Theory of Function Spaces. Birkhäuser Basel.

Tsybakov, A. B. (2008). *Introduction to Nonparametric Estimation*. Springer Science & Business Media.

Urbas, J. (1997). On the second boundary value problem for equations of Monge-Ampère type. *Journal für die reine und angewandte Mathematik*, 1997:115–124.

Vacher, A., Muzellec, B., Bach, F., Vialard, F.-X., and Rudi, A. (2024). Optimal estimation of smooth transport maps with kernel SoS. *SIAM Journal on Mathematics of Data Science*, 6:311–342.

Vacher, A., Muzellec, B., Rudi, A., Bach, F., and Vialard, F.-X. (2021). A dimension-free computational upper-bound for smooth optimal transport estimation. In *Conference on Learning Theory*, pages 4143–4173. PMLR.

van de Geer, S. (2000). *Empirical Processes in M-estimation.* Cambridge UP.

van der Vaart, A. (1998). *Asymptotic Statistics.* Cambridge Series in Statistical and Probabilistic Mathematics. Cambridge University Press, Cambridge, UK ; New York, NY, USA.

van der Vaart, A. W. (2002). Semiparametric statistics. In Bernard, P., editor, *Lectures on Probability Theory and Statistics: École dÉté de Probabilités de Saint-Flour XXIX - 1999*, École d'Éé de Probabilités de Saint-Flour. Springer-Verlag, Berlin Heidelberg.

van der Vaart, A. W. and Wellner, J. A. (1996). *Weak Convergence and Empirical Processes.* Springer.

Vapnik, V. (2013). *The Nature of Statistical Learning Theory.* Springer Science & Business Media.

Vapnik, V. N. and Chervonenkis, A. Y. (1968). The uniform convergence of frequencies of the appearance of events to their probabilities. In *Doklady Akademii Nauk*, volume 181, pages 781–783. Russian Academy of Sciences.

Verdinelli, I. and Wasserman, L. (2019). Hybrid Wasserstein distance and fast distribution clustering. *Electronic Journal of Statistics*, 13:5088–5119.

Verdinelli, I. and Wasserman, L. (2024). Decorrelated variable importance. *Journal of Machine Learning Research*, 25:1–27.

Vershynin, R. (2018). *High-Dimensional Probability: An Introduction with Applications in Data Science.* Cambridge University Press.

Villani, C. (2003). *Topics in Optimal Transportation.* American Mathematical Soc.

Villani, C. (2008). *Optimal Transport: Old and New.* Springer Science & Business Media.

Ville, J. (1939). Étude critique de la notion de collectif. *Bull. Amer. Math. Soc*, 45:824.

Vladimirova, M., Girard, S., Nguyen, H., and Arbel, J. (2020). Sub-Weibull distributions: Generalizing sub-Gaussian and sub-Exponential properties to heavier-tailed distributions. *Stat*, 9.

von Luxburg, U. and Bousquet, O. (2004). Distance-based classification with Lipschitz functions. *Journal of Machine Learning Research*, 5:669–695.

Vovk, V. (2021). Testing randomness online. *Statistical Science*, 36:595–611.

Wainwright, M. J. (2019). *High-Dimensional Statistics: A Non-Asymptotic Viewpoint.* Cambridge University Press.

Wang, Q., Kulkarni, S. R., and Verdú, S. (2005). Divergence estimation of continuous distributions based on data-dependent partitions. *IEEE Transactions on Information Theory*, 51:3064–3074.

Wang, Y. (2019). Convergence rates of latent topic models under relaxed identifiability conditions. *Electronic Journal of Statistics*, 13:37–66.

Waudby-Smith, I. and Ramdas, A. (2024). Estimating means of bounded random variables by betting. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 86:1–27.

Weed, J. (2018). Sharper rates for estimating differential entropy under Gaussian convolutions. *Massachusetts Institute of Technology (MIT), Tech. Rep.*

Weed, J. and Bach, F. (2019). Sharp asymptotic and finite-sample rates of convergence of empirical measures in Wasserstein distance. *Bernoulli*, 25:2620–2648.

Weiss, K., Khoshgoftaar, T. M., and Wang, D. (2016). A survey of transfer learning. *Journal of Big data*, 3:1–40.

Wiechers, H., Eltzner, B., Mardia, K. V., and Huckemann, S. F. (2023). Learning torus PCA-based classification for multiscale RNA correction with application to SARS-CoV-2. *Journal of the Royal Statistical Society Series C: Applied Statistics*, 72:271–293.

Xu, X. (2011). Eigenfunction estimates for Neumann Laplacian and applications to multiplier problems. *Proceedings of the American Mathematical Society*, 139:3583–3599.

Zhang, Q. and Chen, J. (2022). Minimum Wasserstein distance estimator under finite location-scale mixtures. In *Advances and Innovations in Statistics and Data Science*, pages 69–98. Springer.

Zhao, S., Zhou, E., Sabharwal, A., and Ermon, S. (2016). Adaptive concentration inequalities for sequential decision problems. In *Advances in Neural Information Processing Systems 29*, pages 1343–1351.

Zhu, J., Guha, A., Xu, M., Ma, Y., Lei, R., Loffredo, V., Nguyen, X., and Zhao, D. (2021). Functional optimal transport: Mapping estimation and domain adaptation for functional data. *arXiv preprint arXiv:2102.03895.*