THE FAST JOHNSON-LINDENSTRAUSS TRANSFORM

JOHN THICKSTUN

BACKGROUND

The Johnson-Lindenstrauss lemma [1], and more broadly the concept of random projections, is a guiding principle for dimensionality reduction that shows up frequently in theoretical computer science and machine learning. The idea of Johnson-Lindenstrauss is to embed n points from $\mathbb{R}^d$ into $\mathbb{R}^k$ with $k = \Theta(\epsilon^{-2} \log n)$ while distorting distances by at most ϵ.

Lemma. *(Johnson-Lindenstrauss) Let $x_1, \dots, x_n \in \mathbb{R}^d$ and $\epsilon > 0$. If $k > 24\epsilon^{-2} \log n$ then there is some function $A : \mathbb{R}^d \to \mathbb{R}^k$ with*

$$(1) \qquad (1-\epsilon)\|x_i - x_j\|_2 \leq \|Ax_i - Ax_j\|_2 \leq (1+\epsilon)\|x_i - x_j\|_2 \text{ for all } i, j = 1, \dots, n.$$

Proof. See [1]. □

While we omit the proof, we remark that it is constructive. Specifically, A is a linear map consisting of random projections onto subspaces of $\mathbb{R}^d$. These projections can be computed by n matrix multiplications, which take time $O(nkd)$. This is fast enough to make the Johnson-Lindenstrauss transform (JLT) a practical and widespread algorithm for dimensionality reduction, which in turn motivates the desire for an even faster transform. In this report, we will discuss the work of Ailon and Chazelle [2], which speeds up the JLT using a Fourier transform.

Observe that the norms in Johnson-Lindenstrauss are Euclidean. Johnson-Lindenstrauss appears to be a subtle result about ℓ_2; simple counterexamples prevent dimensionality reduction in ℓ_∞, and lower bounds prevent significant reduction in ℓ_1 [3]. The picture is not so gloomy for embeddings. In this report, we will describe a fast method to compute approximate distance-preserving embeddings from $(\mathbb{R}^d, \|\cdot\|_2)$ into $(\mathbb{R}^k, \|\cdot\|_1)$. The same methods extend to the classic ℓ_2 setting that embeds $(\mathbb{R}^d, \|\cdot\|_2)$ into $(\mathbb{R}^k, \|\cdot\|_2)$.

Before moving on, we note that the Johnson-Lindenstrauss reduction is tight [4]. That is, for any n, there exist points $y_1, \dots, y_n \in \mathbb{R}^d$ such that any embedding $A : \mathbb{R}^d \to \mathbb{R}^k$ (linear or otherwise) that approximately preserves distances in the sense of equation (1) must have

$$k = \Omega\left(\frac{\log n}{\epsilon^2}\right).$$

OVERVIEW

Recall that a dense matrix-vector product Ax for $A \in \mathbb{R}^{k\times d}$ and $x \in \mathbb{R}^d$ takes time $O(kd)$. The first idea behind the fast Johnson-Lindenstrauss transform is to replace the dense random matrix A with a sparse matrix P that has the same distance-preserving properties. For sparse P, we can significantly improve the time complexity of the product Px. However, we can no longer make the same adversarial guarantee of the Johnson-Lindenstrauss lemma; if x is sparse then the sparse product Px can easily be small, destroying distance-preservation.

The second idea behind the fast Johnson-Lindenstrauss transform is to preprocess the data x to ensure non-sparsity. This is accomplished by a randomized Fourier transform. The Fourier transform maps sparse vectors to dense vectors. If we additionally randomize the transform with a diagonal sign matrix, then we avoid mapping dense vectors to sparse vectors. The Fourier transform is unitary, so distances are preserved by this preprocessing. And using the FFT, we can compute the preprocessing transform efficiently. With this in mind, we introduce the fast Johnson-Lindenstrauss transform:

Definition. *(FJLT) The* ***Fast Johnson-Lindenstrauss transform*** *is given by the map $\Phi \equiv k^{-1}PHD$, where the terms are defined as follows:*

- *$P : \mathbb{R}^d \to \mathbb{R}^k$ is a sparsified random projection given by the matrix $P \in \mathbb{R}^{k\times d}$ with $P_{ij} \equiv b_{ij}r_{ij}$, where $b_{ij} \sim$ Bernoulli(q) and $r_{ij} \sim \mathcal{N}(0, q^{-1})$ are independent random variables.*
- *$H : \mathbb{R}^d \to \mathbb{R}^d$ is the discrete Fourier transform on the additive group $\mathbb{Z}_2^d$ given by the matrix $H \in \mathbb{R}^{d\times d}$ with $H_{ij} = d^{-1/2}(-1)^{\langle i-1, j-1\rangle}$ where $\langle i, j\rangle$ is the bitwise inner product of the binary representations of i and j.*
- *$D : \mathbb{R}^d \to \mathbb{R}^d$ is a random reflection given by the diagonal matrix $D \in \mathbb{R}^{d\times d}$ where D_{ii} are independent Rademacher random variables.*

Note that the (Walsh-Hadamard) matrix H is dense. The whole purpose of making P sparse is to make the computation $x \mapsto Px$ efficient. But $x \mapsto (PH)Dx$ is a dense matrix-vector multiplication, which as we have discussed requires time $O(kd)$. The point here is that computational complexity is not associative; $x \mapsto (PH)Dx$ is slow but, taking advantage of the FFT and sparse operations, $x \mapsto P(HDx)$ is much faster.

The multiplication Dx is diagonal, so we can compute it in linear time $O(d)$. Using standard fast-Fourier methods, $y \mapsto Hy$ takes time $O(d \log d)$. Using sparse-multiplication methods, $z \mapsto Pz$ takes time $O(|P|)$, the number of non-zero entries of P. So if q (the sparsity parameter) is small then this operation will be efficient. We will show in the next section that we can carefully balance our choice of q so that it is simultaneously large enough to preserve distances and small enough to make the computation $z \mapsto Pz$ efficient. In particular, if ϵ is our distance-distortion tolerance, then we will chose

$$q \equiv \min\left\{\Theta\left(\frac{\log n}{\epsilon d}\right), 1\right\}.$$

ANALYSIS

Using the sparsity parameter q defined previously, we observe that $|P| \sim B(nk, q)$, a binomial distribution. If $n \geq d = \Theta(n)$ then calculating the expectation of the binomial distribution,

$$\mathbb{E}|P| = nkq = O\left(n\frac{\log n}{\epsilon^2}\frac{\log n}{\epsilon d}\right) = O\left(\frac{\log^2 n}{\epsilon^3}\right).$$

Using the standard Chernoff bound for a binomial random variable gives a guarantee that $|P|$ will not be too large. Including preprocessing, with high probability the fast Johnson-Lindenstrauss transform Φ therefore has time complexity

$$O(d \log d + \epsilon^{-3} \log^2 n).$$

In the following lemma, we will show that the FJLT approximately preserves norms with constant probability. We can convert the constant probability result into a high probability result by repeated re-construction of the transform; to achieve the desired level of distortion with probability $1 - \delta$ requires $O(\log(\delta^{-1}))$ constructions of Φ, which only changes the time-complexity of the FJLT by a multiplicative log factor. After proving the lemma, we will show how to extend the result for norms to the desired result for metrics.

Lemma 1. *Let $x_1, \ldots, x_n \in \mathbb{R}^d$ and $\epsilon < 1$. For $i = 1, \ldots, n$, with non-vanishing probability,*

$$(1-\epsilon)\sqrt{2\pi^{-1}}\|x_i\|_2 \leq \|\Phi x_i\|_1 \leq (1+\epsilon)\sqrt{2\pi^{-1}}\|x_i\|_2.$$

Proof. This proof proceeds as follows. First, we show that after pre-processing with HD, data points are unlikely to concentrate at particular coordinates. We then appeal to Lemma 2 from [2] to prove the result in expectation. Finally, we show that the transformed data points concentrate around their mean.

If $x \in \mathbb{R}^d$ and $\|x\|_2 = 1$ then we will show that $\|HDx\|_\infty$ is likely to be near $1/\sqrt{d}$. Specifically, we define $u \equiv HDx$ and construct a Chernoff bound on the size of each coordinate; wolog we will argue in the case of the first coordinate that

$$P\left(|u_1| \geq \Theta\left(\frac{\sqrt{\log n}}{\sqrt{d}}\right)\right) \leq \frac{1}{20nd}.$$

By construction of D, u is symmetrically distributed, and by Markov's inequality,

$$P(|u_1| \geq s) = 2P(u_1 \geq s) = 2P(e^{sdu_1} \geq e^{s^2 d}) \leq 2e^{-s^2 d}\mathbb{E}e^{sdu_1}.$$

We proceed by the method of Laplace transforms. Observe that $u_1 = \sum_{i=1}^d a_i x_i$ with a_i are uniform i.i.d. $\pm 1/\sqrt{d}$ random variables, so

$$\mathbb{E}e^{sdu_1} = \prod_{i=1}^d \mathbb{E}e^{sda_i x_i} = \prod_{i=1}^d \cosh(s\sqrt{d}x_i).$$

Expanding the hyperbolic cosine as an infinite product and recalling that $e^x \geq 1 + x$,

$$\cosh(t) = \prod_{j=1}^\infty \left(1 + \frac{4t^2}{\pi^2(2j-1)^2}\right) \leq \exp\left(\sum_{j=1}^\infty \frac{4t^2}{\pi^2(2j-1)^2}\right) = \exp(t^2/2).$$

This allows us to bound the moment generating function:

$$\mathbb{E}e^{sdu_1} = \prod_{i=1}^{d} \cosh(s\sqrt{d}x_i) \leq \prod_{i=1}^{d} e^{s^2 d x_i^2/2} = e^{s^2 d\|x\|_2^2/2} = e^{s^2 d/2}.$$

Choose c so that $n^c > 20$; if $s \equiv d^{-1/2}\sqrt{2(2+c)\log n} = \Theta(d^{-1/2}\sqrt{\log n})$ then

$$P(|u_1| \geq s) \leq 2e^{-s^2 d/2} = 2e^{-(2+c)\log n} = \frac{2}{n^{2+c}} \leq \frac{1}{20nd}.$$

And by the union bound,

$$P\left(\max_{x_1,\dots,x_n} \|HDx\|_\infty > O(\sqrt{\log n}/\sqrt{d})\right) \leq \sum_{i=1}^{n}\sum_{j=1}^{d} P\left(|(HDx_i)_j| \geq s\right) \leq \frac{1}{20}.$$

We now proceed under the assumption that $\|u\|_\infty \leq s$. First, observe that $\|u\|_2 = \|x\|_2$ because H (random reflections) and D (a Fourier transform) are isometries. Define

$$y \equiv Pu = k\Phi x.$$

By definition of Φ, we can write

$$y_1 = \sum_{j=1}^{d} r_{1j} b_{1j} u_j.$$

Analysis of the other coordinates of y is identical, so we will restrict our analysis to y_1 and write r_j and b_j (omitting the constant first index). Recall that $r_j \sim \mathcal{N}(0, q^{-1})$ and $\sum_{j=1}^{d} r_j\alpha_j \sim \mathcal{N}\left(0, q^{-1}\sum_{j=1}^{d}\alpha_j^2\right)$. Therefore the distribution of y_1 is

$$\sum_{j=1}^{d} r_j(b_j u_j) \sim \mathcal{N}\left(0, q^{-1}\sum_{j=1}^{d} b_j^2 u_j^2\right) = \mathcal{N}\left(0, q^{-1}\sum_{j=1}^{d} b_j u_j^2\right).$$

Define $Z \equiv \sum_{j=1}^{d} b_j u_j^2$ and recall the expectation of the half-normal distribution to see that

$$\mathbb{E}|y_1| = \mathbb{E}\left[\mathbb{E}\left[|y_1||Z\right]\right] = \mathbb{E}\left[\sqrt{\frac{2q^{-1}Z}{\pi}}\right] = \sqrt{\frac{2}{q\pi}}\mathbb{E}\sqrt{Z}. \tag{2}$$

We now quote a result from [2]:

Lemma 2. *(Ailon and Chazelle) For any $t > 1$, $\mathbb{E}Z^t = O(qt)^t$ and*

$$(1-\epsilon)\sqrt{q} \leq \mathbb{E}\sqrt{Z} \leq \sqrt{q}.$$

It follows that from the second observation of Lemma 2 that

$$(1-\epsilon)\sqrt{2\pi^{-1}} \leq \mathbb{E}|y_1| \leq \sqrt{2\pi^{-1}}.$$

By symmetry $\mathbb{E}\|y\|_1 = k\mathbb{E}|y_1|$, which gives us our result in expectation; we now need to show that $\|y\|_1$ concentrates around its mean.

We will construct a Chernoff bound to control the dispersion of $|y_1|$. Let $U = y_1/\sqrt{q^{-1}Z}$. Observe that $U \sim \mathcal{N}(0,1)$ given Z and

$$\mathbb{E}|y_1|^t = \mathbb{E}\left[\mathbb{E}\left[|y_1|^t|Z\right]\right] = \mathbb{E}\left[(q^{-1}Z)^{t/2}\mathbb{E}\left[|U|^t|Z\right]\right] = \mathbb{E}\left[(q^{-1}Z)^{t/2}\right]\mathbb{E}\left[|U|^t|Z\right].$$

Analysis of the half-normal distribution shows that $\mathbb{E}\left[|U|^t|Z\right] = O(t)^{t/2}$ and combining this with Lemma 2 (the first observation)

$$\mathbb{E}|y_1|^t = q^{-t/2}O(qt/2)^{t/2}O(t)^{t/2} = O(t)^t.$$

Taylor-expanding the moment generating function, we have

$$\mathbb{E}e^{\lambda|y_1|} = 1 + \lambda\mathbb{E}|y_1| + \sum_{t>1}\mathbb{E}|y_1|^t\lambda^t/t! \leq 1 + \lambda\mathbb{E}|y_1| + \sum_{t>1}O(t)^t\lambda^t/t!.$$

By Stirling's approximation, if $\lambda < \lambda_0 \approx e^{-1}$ then the mgf converges and

$$\mathbb{E}e^{\lambda|y_1|} \leq 1 + \lambda\mathbb{E}|y_1| + \sum_{t>1}O(1)e^t\lambda^t = 1 + \lambda\mathbb{E}|y_1| + O(\lambda^2) = e^{\lambda\mathbb{E}|y_1| + O(\lambda^2)}.$$

Because the components of y are i.i.d.,

$$\mathbb{E}e^{\lambda\|y\|_1} = \left(\mathbb{E}e^{\lambda|y_1|}\right)^k = e^{\lambda\mathbb{E}\|y\|_1 + O(\lambda^2 k)}.$$

Therefore by Markov's inequality, using the expectation of $\|y\|_1$ computed earlier,

$$P\left(\|y\|_1 \geq (1+\epsilon)\mathbb{E}\|y\|_1\right) \leq e^{-\lambda(1+\epsilon)\mathbb{E}\|y\|_1}\mathbb{E}e^{\lambda\|y\|_1} \leq e^{-\lambda\epsilon\mathbb{E}\|y\|_1 + O(\lambda^2 k)} \leq e^{-\Omega(\lambda^2 k)}.$$

A similar argument bounds the left tail. Setting an appropriate $\lambda = \Theta(\epsilon) < \lambda_0$ together with $k = \Theta(c\epsilon^{-2}\log n)$ (with $n^c > 20$) then $\|y\|_1$ deviates from its mean by more than ϵ with probability less than

$$e^{-\Omega(\epsilon^2 k)} = e^{-c\log n} = \frac{1}{n^c} \leq \frac{1}{20}. \qquad \square$$

Recall from earlier that we can upgrade the result of Lemma 1 to a $1-\delta$ high probability statement; let $\delta = n^{-3}$. Define $y_{ij} = x_i - x_j$; there are n^2 of these vectors so applying the union bound together with Lemma 1, with probability $1/n$, for all $i, j = 1, \ldots, n$,

$$(1-\epsilon)\sqrt{2\pi^{-1}}\|x_i - x_j\|_2 \leq \|\Phi x_i - \Phi x_j\|_1 \leq (1+\epsilon)\sqrt{2\pi^{-1}}\|x_i - x_j\|_2. \tag{3}$$

This result using the efficient FJLT Φ is directly comparable to equation (1). We only lose a $\sqrt{2\pi^{-1}}$ factor due to the mismatch between the ℓ_1 and ℓ_2 norms (this factor disappears in the $\ell^2 \mapsto \ell^2$ version of the FJLT).

References

[1] William B Johnson and Joram Lindenstrauss. Extensions of Lipschitz mappings into a Hilbert space. Contemporary mathematics (1984).

[2] Nir Ailon and Bernard Chazelle. The Fast Johnson-Lindenstrauss Transform and Approximate Nearest Neighbors. SCIAM Journal on Computing (2009).

[3] James R. Lee and Assaf Naor. Embedding the diamond graph in L_p and dimension reduction in L_1. Geometric and Functional Analysis (2004).

[4] Kasper Green Larsen and Jelani Nelson. Optimality of the Johnson-Lindenstrauss Lemma. arXiv preprint 1609.02094 (2016).