Some Useful Background for Talk on the Fast Johnson-Lindenstrauss Transform

Nir Ailon*

May 22, 2007

This writeup includes very basic background material for the talk on the *Fast Johnson Lindenstrauss Transform* on May 28th, 2007 by the author at the Summer School on Algorithmic Data Analysis in Helsinki, Finland. The talk is about the results in [2], but will be based on a new more recent proof (together with Edo Liberty), using more modern tools, such as probability in Banach spaces and error-correcting codes. These tools also give stronger results. A written version of the new proof will be available for download as a technical report shortly. The information here touches the surface of deep, classic theories, but should be useful for those who are interested in a quick review. The talk will "fuse" the seemingly disparate ideas here into an algorithm. Some exercises are included.

1 Dimension Reduction

In many applications, we are given as input points in high dimension and wish to reduce the dimension while (approximately, with high probability) preserving certain properties. Such a procedure can be used to save space or time, or to remove redundancy from data. The above definition is extremely loose and can encompass buzz words such as "Principal Component Analysis (PCA)", "sparse approximation" and even "compression". I will talk about dimension reduction from the perspective of *algorithmic metric embedding.* This means that the property we preserve is the metric on the points, and we also care about the computational resources required for the reduction.

There are two metrics in question: the domain (original) metric, which will always be the ℓ_2 (Euclidean) metric in $\mathbf{R}^d$ here and the range (reduced) metric, which will either be ℓ_1 ("Manhattan distance") or ℓ_2. We remind that the Euclidean distance between two vectors $u, v \in \mathbf{R}^d$ is $\|u-v\|_2 = (\sum(u_i - v_i)^2)^{1/2}$ and the Manhattan distance is $\sum |u_i - v_i|$.

Given a set of n points V in $\mathbf{R}^d$ for some large dimension d, the goal to apply a random map Φ from $\mathbf{R}^d$ to $\mathbf{R}^k$ (for $k \ll d$) so that with probability 2/3 (say), simultaneously for all pairs of points $u, v \in V$,

$$(1-\varepsilon)\|u-v\|_2 \leq \|\Phi u - \Phi v\|_p \leq (1+\varepsilon)\|u-v\|_2 \ ,$$

where $p = 1$ or $p = 2$. We will use $\|\Phi u - \Phi v\|_p \approx_\varepsilon \|u-v\|_2$ as shorthand for the last sandwich inequality.

A classic result by Johnson and Lindenstrauss [6] and certain simplifications [1, 3, 5, 4] show that if $k = O(\varepsilon^{-2} \log n)$, then a "random transformation" does the trick. The main argument in

[*] Institute for Advanced Study, Princeton NJ. Email: `nailon@math.ias.edu`

all these proof is: Show using a measure concentration principle that

$$\forall x \in \mathbf{R}^d,\ \Pr[\|\Phi x\|_p \not\approx_\varepsilon \|x\|_2] = e^{-\Omega(k\varepsilon^2)}\ . \tag{1}$$

Now plug in k as above and apply a union bound over all $\binom{n}{2}$ possibilities $x = u - v$ for $u, v \in V$.

It makes sense to abstract this and say that a distribution on $\mathbf{R}^d \to \mathbf{R}^k$ linear transformations has the "Johnson-Lindenstrauss" property if (1) holds for Φ drawn from this distribution. *Note the quantification:* For any fixed (adversarial) x, $\|\Phi x\|_p \approx_\varepsilon \|x\|$ with high probability. By this we actually mean that a *family* of distributions for increasingly large k, d has this property with the same constant hiding in the Ω-notation. Once we have this definition, and since we know already it is not an empty one, it makes sense to ask what is the best distribution satisfying the J-L property in terms of computational resources? More precisely, we try to optimize

- the time it takes to reduce a vector, and
- the number of random bits used.

(The latter may seem not very important for applications but reducing this quantity can in fact reduce the amount of space required when applying the dimension reduction to data streams.) For example, the constructions in [1, 3, 5, 4] require $O(kd)$ time and random bits. The construction in [2] requires $O(\max\{d \log d, k^3\})$ time for $p = 2$ and $O(\max\{d \log d, k^2\})$ for $p = 1$. A new proof that will be given in the talk offers better results. We review some relevant background in this writeup.

2 Finite Dimensional Real Banach Spaces

A Banach space is a normed vector space satisfying a certain topological property which we will not really need in the talk. Here we only consider finite dimensional real Banach spaces and some standard norms. Some definitions and facts:

- For $1 \le p < \infty$ the space ℓ_p^d is $\mathbf{R}^d$ equipped with the norm $\|x\| = \|x\|_p = \left(\sum |x_i|^p\right)^{1/p}$. For $p = \infty$, ℓ_p^d is $\mathbf{R}^d$ equipped with the norm $\|x\| = \|x\|_\infty = \max |x_i|$. We call p the *norm index* or *norm exponent*.
- For norm index p the *dual norm* q is defined by $1/q + 1/p = 1$.
- If p, q are dual norm indices, then for all $x, y \in \mathbf{R}^d$,

  $$\langle x, y \rangle = \sum x_i y_i \le \|x\|_p \|y\|_q\ .$$

  This is called the Hölder inequality. For $p = q = 2$, this is commonly known as Cauchy-Schwartz inequality.
- If p, q are dual norm indices, then for $x \in \ell_p^d$

  $$\|x\|_p = \sup_{y \in \ell_q^d, \|y\|_q \le 1} \langle x, y \rangle\ .$$

- For a matrix $A \in \mathbf{R}^{k\times d}$ and two norm indices p_1, p_2 we define the operator norm $\|A\|_{p_1\to p_2}$ as

  $$\sup_{x\in\ell_{p_1}^d, \|x\|\le 1} \|Ax\|_{p_2} \ .$$

  Equivalently, using the above,

  $$\|A\|_{p_1\to p_2} = \sup_{x\in\ell_{p_1}^d, \|x\|\le 1} \sup_{y\in\ell_{q_2}^d, \|y\|\le 1} y^T Ax \ ,$$

  where q_2 is the dual norm to p_2.

  Exercise: Convince yourself that $\|A\|_{p_1\to p_2} = \|A^T\|_{q_2\to q_1}$, where q_1, q_2 are dual to p_1, p_2, respectively.

- **Riesz-Thorin interpolation theorem**: Let p_1, r_1, p_2, r_2 be some norm indices. Let $A \in \mathbf{R}^{k\times d}$, and assume $\|A\|_{p_1\to r_1} \le C_1$ and $\|A\|_{p_2\to r_2} \le C_2$. Let $0 \le \lambda \le 1$, and let p, r be such that $\frac{1}{p} = \lambda\frac{1}{p_1} + (1-\lambda)\frac{1}{p}_2$ and $\frac{1}{r} = \lambda\frac{1}{r_1} + (1-\lambda)\frac{1}{r}_2$. (geometrically, the point $(1/p, 1/r)$ is on the line connecting $(1/p_1, 1/r_1)$ and $(1/p_2, 1/r_2)$). Then $\|A\|_{p\to r} \le C_1^\lambda C_2^{1-\lambda}$.

3 The Walsh-Hadamard Matrix

It is sometimes useful to assume that the coordinates of $\mathbf{R}^d$ are structured. One commonly used structure is the discrete line. Often it is useful to assume that $d = 2^t$ for some integer t and coordinate i is a t-dimensional vector over $\mathbb{F}_2$ (the field with two elements $0, 1$) representing the number in binary.

The Walsh-Hadamard matrix $H_d \in \mathbf{R}^{d\times d}$ is defined by

$$H_d(i,j) = d^{-1/2}(-1)^{\langle i,j\rangle} \ ,$$

where $\langle i,j\rangle$ is the scalar product of i, j over $\mathbb{F}_2^t$. It is a type of Fourier transform matrix.
Exercise: Prove that H_d can be recursively written as

$$H_d = \frac{1}{\sqrt{2}}\begin{pmatrix} H_{d/2} & H_{d/2} \\ H_{d/2} & -H_{d/2}\end{pmatrix} \ .$$

Using this, convince yourself that $x \mapsto H_d x$ can be computed in time $O(d\log d)$ for a vector $x \in \mathbf{R}^d$.
Exercise: Prove that H_d is an orthogonal matrix (the columns have unit Euclidean length and are orthogonal to each other).
Exercise: Let $1 \le p \le 2$ be a norm index, and let q be its dual. Prove that $\|H_d\|_{p\to q} \le d^{1/2-1/p}$ (Hint: find $\|H_d\|_{1\to\infty}$ and use Riesz-Thorin). This is often known as the **Hausdorff-Young** theorem.

4 Measure Concentration on the Hypercube

Let f be a function from ℓ_2^d to $\mathbf{R}$ that is C-Lipschitz (this means that for any $x, y \in \ell_2^d$, $|f(x)-f(y)| \le C\|x-y\|_2$) and *convex*.

Example: $f(x) = \|Ax\|_p$ for some matrix $A \in \mathbf{R}^{k\times d}$. Then f is convex and $(\|A\|_{2\to p})$-Lipschitz.

Now let $X_1, \ldots, X_d$ be d independent $\{\pm 1\}$ random variables. Let M_f be a median of $f(X_1, \ldots, X_d)$. That is, M_f is a number such that both $\Pr[f(X_1, \ldots, X_d) \geq M_f] \geq 1/2$ and $\Pr[f(X_1, \ldots, X_d) \leq M_f] \geq 1/2$. Then

$$\Pr[|f(X_1, \ldots, X_d) - M_f| > t] \leq 4\exp(-t^2/8C^2) \ .$$

This was proved by Talagrand (see proof in [7]) and is a powerful generalization of many other measure concentration theorems.

5 Error Correcting Codes

Error correcting codes are combinatorial objects with many nice properties useful in engineering and also in theoretical computer science. A good introduction can be found in [8]. Another great resource is Madhu Sudan's lecture notes http://theory.lcs.mit.edu/˜madhu/FT01/. Here we concentrate on the definition of linear binary error correcting codes. We will not discuss constructions.

The field of interest is $\mathbb{F}_2$. A $[k, m, D]$-code is a linear subspace of $\mathbb{F}_2^k$ of dimension $m \leq k$ with the property that for any nonzero $x \in \mathbb{F}_2^k$, $\Delta(x) \geq D$, where $\Delta(x)$ is the Hamming weight of x (number of nonzeros).

If V is a $[k, m, D]$-code, then the dual code $V^\perp$ is the space of all vectors $y \in \mathbb{F}_2^k$ such that the scalar product $\langle x, y\rangle = 0$ for all $x \in V$. By standard linear algebra, $\dim V^\perp = k - m$.
Exercise:

1. A set of vectors $U \subseteq \mathbb{F}_2^k$ is s-wise independent if for any set $i_1 < i_2 \cdots < i_s \in [k]$ of coordinates, the number of vectors $x \in U$ such that $(x_{i_1}, \ldots, x_{i_s}) = y$ is exactly $|U|/2^s$, for any $y \in \mathbb{F}_2^s$. Let $V^\perp \subseteq \mathbb{F}_2^k$ be the dual of a $[k, m, D]$-code. Prove that $V^\perp$ is $(D-1)$-wise independent.

2. Let $V^\perp$ be as above, and assume in addition that $D \geq 3$ and odd. We transform $V^\perp$ into a matrix A of size $2^{k-m} \times k$ by writing each codeword $x \in V^\perp$ as a row of A, turning 0's to $(+1)$'s and 1's to (-1)'s (note that $V^\perp$ contains exactly 2^{k-m} vectors). Prove that $\|A\|_{2\to(D-1)} \leq c2^{(k-m)/(D-1)}$ where c is a constant that depends on D *only* (and not on k, m). (Hint: Consider the random variable $(y \cdot x)^{D-1}$ for a random row y of A and a fixed $x \in \ell_2^d$). Find the best bound you can on $\|A\|_{2\to p}$ for $2 \leq p \leq D-1$ (Hint: Find $\|A\|_{2\to 2}$ exactly and then apply Riesz-Thorin).

3. Let A be as above. Prove that if we permute the rows of A then (up to a constant factor) the columns of A are a subset of the columns of $H_{2^{k-m}}$. (Hint: We only need the fact that $V^\perp$ is a vector space.) Conclude that $x \mapsto A^T x$ can be "efficiently" computed for any $x \in \mathbf{R}^{2^{k-m}}$.

References

[1] D. Achlioptas. Database-friendly random projections: Johnson-Lindenstrauss with binary coins. *J. Comput. Syst. Sci.*, 66(4):671–687, 2003.

[2] N. Ailon and B. Chazelle. Approximate nearest neighbors and the fast Johnson-Lindenstrauss transform. In *Proceedings of the 38st Annual Symposium on the Theory of Compututing (STOC)*, pages 557–563, Seattle, WA, 2006.

[3] S. DasGupta and A. Gupta. An elementary proof of the Johnson-Lindenstrauss lemma. *Technical Report, UC Berkeley*, 99-006, 1999.

[4] P. Frankl and H. Maehara. The Johnson-Lindenstrauss lemma and the sphericity of some graphs. *Journal of Combinatorial Theory Series A*, 44:355–362, 1987.

[5] P. Indyk and R. Motwani. Approximate nearest neighbors: Towards removing the curse of dimensionality. In *Proceedings of the 30th Annual ACM Symposium on Theory of Computing (STOC)*, pages 604–613, 1998.

[6] W. B. Johnson and J. Lindenstrauss. Extensions of Lipschitz mappings into a Hilbert space. *Contemporary Mathematics*, 26:189–206, 1984.

[7] M. Ledoux and M. Talagrand. *Probability in Banach Spaces: Isoperimetry and Processes.* Springer-Verlag, 1991.

[8] F. MacWilliams and N. Sloane. *The Theory of Error Correcting Codes.* North-Holland, 1983.