INTERNATIONAL COMPUTER SCIENCE INSTITUTE

1947 Center St. • Suite 600 • Berkeley, California 94704-1198 • (510) 643-9153 • FAX (510) 643-7684

An elementary proof of the Johnson-Lindenstrauss Lemma

Sanjoy Dasgupta * **Anupam Gupta** †

TR-99-006

Abstract

The Johnson-Lindenstrauss lemma shows that a set of n points in high dimensional Euclidean space can be mapped down into an $O(\log n/\epsilon^2)$ dimensional Euclidean space such that the distance between any two points changes by only a factor of $(1 \pm \epsilon)$. In this note, we prove this lemma using elementary probabilistic techniques.

*Computer Science Division, UC Berkeley. Email: dasgupta@cs.berkeley.edu.

†Computer Science Division, UC Berkeley. Email: angup@cs.berkeley.edu. Supported by NSF grant CCR-9505448.

1 Introduction

Johnson and Lindenstrauss [6] proved a fundamental result, which said that any n point set in Euclidean space could be embedded in $O(\log n/\epsilon^2)$ dimensions without distorting the distances between any pair of points by more than a factor of $(1 \pm \epsilon)$, for any $0 < \epsilon < 1$. Recently, this lemma has found several applications, including Lipschitz embeddings of graphs into normed spaces [7] and searching for approximate nearest neighbours [5].

The original proof of Johnson and Lindenstrauss was much simplified by Frankl and Maehara [2, 3], using geometric insights and refined approximation techniques. The proof given in this note uses elementary probabilistic techniques to obtain essentially the same results. Independently, Indyk and Motwani [5] have obtained a simple proof of a slightly weaker version of our structural lemma (2.2).

2 The Johnson-Lindenstrauss Lemma

Theorem 2.1 (Johnson-Lindenstrauss lemma) *For any* $0 < \epsilon < 1$ *and any integer* n*, let* k *be a positive integer such that*

$$k \geq 4(\epsilon^2/2 - \epsilon^3/3)^{-1} \ln n. \tag{1}$$

Then for any set V *of* n *points in* R^d*, there is a map* $f : \mathrm{R}^d \to \mathrm{R}^k$ *such that for all* $u, v \in V$*,*

$$(1-\epsilon)||u-v||^2 \;\leq\; ||f(u)-f(v)||^2 \;\leq\; (1+\epsilon)||u-v||^2.$$

Further this map can be found in randomized polynomial time.

The theorem is proved by showing that the squared length of a random vector is sharply concentrated about its mean when the vector is projected onto a random k dimensional subspace, and is not distorted by more than $(1 \pm \epsilon)$ with probability $O(1/n^2)$. Applying the trivial union bound then gives the theorem.

Hence the aim is to estimate the length of a unit vector in R^d when it is projected onto a random k-dimensional subspace. However, this length has the same distribution as the length of a random unit vector projected down onto a fixed k-dimensional subspace. Here we take this subspace to be the space spanned by the first k coordinate vectors, for simplicity.

Let $X_1, \ldots, X_d$ be d independent $N(0,1)$ random variables, and let $Y = \frac{1}{||X||}(X_1, \ldots, X_d)$. It is simple to see that Y is a point chosen uniformly at random from the surface of the d-dimensional sphere S^{d-1}. Let the vector $Z \in \mathrm{R}^k$ be the projection of Y onto its first k coordinates, and let $L = ||Z||^2$. Clearly the expected length $\mu = EL = k/d$. We want to show that L is also fairly tightly concentrated around μ.

Lemma 2.2 *Let* $k < d$*. Then*

2.2a. If $\beta < 1$ *then*

$$\Pr[L \leq \beta k/d] \;\leq\; \beta^{k/2}\left(1 + \frac{(1-\beta)k}{(d-k)}\right)^{(d-k)/2} \;\leq\; \exp(\frac{k}{2}(1-\beta+\ln\beta)).$$

2.2b. If $\beta > 1$ then

$$\Pr[L \geq \beta k/d] \leq \beta^{k/2}\left(1 + \frac{(1-\beta)k}{(d-k)}\right)^{(d-k)/2} \leq \exp(\frac{k}{2}(1-\beta+\ln\beta)).$$

The proofs are similar to those for the Chernoff-Hoeffding bounds [1, 4] and are given in Section 3. ∎

Proof: (Theorem 2.1)
If $d \leq k$, then the theorem is trivial. Else we take a random k-dimensional subspace S, and let v_i' be the projection of vertex $v_i \in V$ into S. Then setting $L = ||v_i' - v_j'||^2$ and $\mu = (k/d)||v_i - v_j||^2$, and applying lemma (2.2a), we get that

$$\begin{aligned}\Pr[L \leq (1-\epsilon)\mu] &\leq \exp\left(\frac{k}{2}(1-(1-\epsilon)+\ln(1-\epsilon))\right)\\ &\leq \exp\left(\frac{k}{2}(\epsilon-(\epsilon+\epsilon^2/2))\right) = \exp\left(-\frac{k\epsilon^2}{4}\right)\\ &\leq \exp(-2\ln n) = 1/n^2,\end{aligned}$$

where, in the second line, we have used the inequality $\ln(1-x) \leq -x - x^2/2$, valid for all $x \geq 0$.

Similarly, we can apply lemma (2.2b) and the inequality $\ln(1+x) \leq x - x^2/2 + x^3/3$ (which is valid for all $x > 0$) to get

$$\begin{aligned}\Pr[L \geq (1+\epsilon)\mu] &\leq \exp\left(\frac{k}{2}(1-(1+\epsilon)+\ln(1+\epsilon))\right)\\ &\leq \exp\left(\frac{k}{2}(-\epsilon+(\epsilon-\epsilon^2/2+\epsilon^3/3))\right) = \exp\left(-\frac{k(\epsilon^2/2-\epsilon^3/3)}{2}\right)\\ &\leq \exp(-2\ln n) = 1/n^2,\end{aligned}$$

Now we can choose the map $f(v_i) = (\sqrt{n/k})v_i'$. By the above calculation, the chance that for some fixed pair i, j, the distortion $||f(v_i) - f(v_j)||^2/||v_i - v_j||^2$ does not lie in the range $[(1-\epsilon), (1+\epsilon)]$ is at most $2/n^2$. Using the trivial union bound, the chance that some pair of vertices suffers a large distortion is at most $\binom{n}{2} \times 2/n^2 = (1 - \frac{1}{n})$. Hence f has the desired properties with probability at least $1/n$. Repeating this projection $O(n)$ times can boost the success probability to any desired constant, giving us the claimed randomized polynomial time algorithm. ∎

3 An Application: Embedding into arbitrary dimensions

Let (X, ρ) be a finite metric. For a map $f : X \to \mathrm{R}^k$, we define

$$||f||_L = \max_{x,y\in X}\frac{||f(x)-f(y)||}{\rho(x,y)} \text{ and } ||f^{-1}||_L = \max_{x,y\in X}\frac{\rho(x,y)}{||f(x)-f(y)||}.$$

Now the Lipschitz distortion of the map f is defined to be $||f||_L \cdot ||f^{-1}||_L$.

As a simple corollary of our structure lemma 2.2, we can deduce that any weighted graph can be embedded into $k \leq C \log n$ dimensions with only $O(n^{2/k}(\log n)^{3/2}/\sqrt{k})$ distortion, and that such a map can be found in randomized polynomial time. This result was earlier proved by Matoušek who had used the same projection technique, but he proved the distortion bound from first principles.

To perform the embedding, we first embed the graph into ℓ_2 with $O(\log n)$ distortion [7], and then project it down to k dimensions using the technique above. If we can show that a unit vector projected down onto k dimensions has $D_2 k/n \leq L \leq D_1 k/n$ with probability $1 - 1/n^2$, then using the union bound over all $\binom{n}{2}$ pairwise distances, we would have a projection with $\sqrt{D_1/D_2}$ distortion with probability at least $1/n$. Note that we obtain a square root because L is the square of the Euclidean length, while the distortion is defined in terms of the Euclidean length. Taking into account the distortion in the first step, the total distortion would be $O(\log n \sqrt{D_1/D_2})$.

Applying Lemma (2.2a) with β being $D_2 = (\mathrm{e}n^{4/k})^{-1}$ gives us that $\Pr[L \leq D_2 k/n] \leq 1/n^2$. Further, taking β to be $D_1 = (7 \max\{1, C\} \ln n)/k$ in Lemma (2.2b) gives $\Pr[L \geq D_1 k/n] \leq 1/n^2$. The proofs of these facts involve routine calculations using the fact that $k \leq C \log n$. Hence, with probability $O(1/n^2)$, $D_2 k/n \leq L \leq D_1 k/n$. Thus, with probability $O(1/n)$, the distortion due to the projection itself is

$$\sqrt{D_1/D_2} = cn^{2/k}((\ln n)/k)^{1/2} = c \exp\{2(\ln n)/k\}((\ln n)/k), \tag{2}$$

where $c = \sqrt{7\mathrm{e}\max\{1, C\}}$; and the total distortion is $c \log n \sqrt{D_1/D_2} = O(n^{2/k}(\log n)^{3/2}/\sqrt{k})$.

In the same paper [8], Matoušek showed that for every integer l, there exists a set of n points in R^{2l+1} which requires a distortion of $\Omega(n^{1/l})$ to embed into R^{2l}. In this sense, the projection technique (and the analysis) is almost optimal.

4 Proofs of tail bounds

Proof of Lemma (2.2a):

We use the fact that if $X \sim N(0,1)$, then $E[e^{tX^2}] = 1/\sqrt{1-2t}$, for $-\infty < t < \frac{1}{2}$. We now prove that

$$\Pr[d(X_1^2 + \cdots + X_k^2) \leq k\beta(X_1^2 + \cdots + X_d^2)] \leq \beta^{k/2}\left(1 + \frac{k(1-\beta)}{d-k}\right)^{(d-k)/2} \tag{3}$$

Note that this is just another way of stating lemma 2.2a

$$\begin{aligned}
& \Pr[d(X_1^2 + \cdots + X_k^2) \leq k\beta(X_1^2 + \cdots + X_d^2)] \\
= \; & \Pr[k\beta(X_1^2 + \cdots + X_d^2) - d(X_1^2 + \cdots + X_k^2) \geq 0] \\
= \; & \Pr[\exp\left\{t(k\beta(X_1^2 + \cdots + X_d^2) - d(X_1^2 + \cdots + X_k^2))\right\} \geq 1] && \text{(for } t > 0\text{)} \\
\leq \; & \mathrm{E}[\exp\left\{t(k\beta(X_1^2 + \cdots + X_d^2) - d(X_1^2 + \cdots + X_k^2))\right\}] && \text{(by Markov's inequality)}
\end{aligned}$$

$$
\begin{aligned}
&= \mathrm{E}\left[\exp\left\{tk\beta X^2\right\}\right]^{(d-k)}\mathrm{E}\left[\exp\left\{t(k\beta-d)X^2\right\}\right]^k \qquad \text{(where } X \sim N(0,1)\text{)}\\
&= (1-2tk\beta)^{-(d-k)/2}(1-2t(k\beta-d))^{-k/2} = g(t).
\end{aligned}
$$

The last line of the derivation gives us the additional constraints that $tk\beta < \frac{1}{2}$ and $t(k\beta - d) < \frac{1}{2}$. The latter constraint is subsumed by the former (since $t \geq 0$), and so $0 < t < 1/2k\beta$. Now to minimize $g(t)$, we maximize

$$f(t) = (1-2tk\beta)^{(d-k)}(1-2t(k\beta-d))^k$$

in the interval $0 < t < 1/2k\beta$. Differentiating f, we get that the maximum is achieved at

$$t_0 = \frac{(1-\beta)}{2\beta(d-k\beta)}.$$

which lies in the permitted range $(0, 1/2k\beta)$. Hence we have

$$f(t_0) = \left(\frac{d-k}{d-k\beta}\right)^{d-k}\left(\frac{1}{\beta}\right)^k$$

and the fact that $g(t_0) = 1/\sqrt{f(t_0)}$ proves the inequality (3). ∎

Proof of Lemma (2.2b):

The proof is almost exactly the same as that of lemma (2.2a). The same calculations show

$$
\begin{aligned}
&\Pr[d(X_1^2+\cdots+X_k^2) \geq k\beta(X_1^2+\cdots+X_d^2)]\\
=\ &(1+2tk\beta)^{-(d-k)/2}(1+2t(k\beta-d))^{-k/2} = g(-t).
\end{aligned}
$$

for $0 < t < 1/2(d-k\beta)$. But this will be minimized at $-t_0$, where t_0 was as defined in the previous proof. This does lie in the desired range $(0, 1/2(d-k\beta))$ for $\beta > 1$, which gives us that

$$\Pr[d(X_1^2+\cdots+X_k^2) \geq k\beta(X_1^2+\cdots+X_d^2)] \leq \beta^{k/2}\left(1+\frac{k(1-\beta)}{d-k}\right)^{(d-k)/2}.$$

∎

References

[1] Chernoff, H. Asymptotic efficiency for tests based on the sum of observations. *Ann. Math. Stat.* **23**, 1952, pp. 493–507.

[2] Frankl, P. and Maehara, H. The Johnson-Lindenstrauss lemma and the sphericity of some graphs. *J. Combin. Theory Ser. B* **44**(3), 1988, pp. 355–362.

[3] Frankl, P. and Maehara, H. Some geometric applications of the beta distribution. *Ann. Inst. Stat. Math.* **42**(3), 1990, pp. 463–474.

[4] Hoeffding, W. Probability for sums of bounded random variables. *J. American Stat. Assoc.* **58**, 1963, pp. 13–30.

[5] Indyk, P. and Motwani, R. Approximate Nearest Neighbors: Towards Removing the Curse of Dimensionality. *Proc. 30th Symposium on Theory of Computing*, 1998, pp. 604–613.

[6] Johnson, W. and Lindenstrauss, J. Extensions of Lipschitz maps into a Hilbert space. *Contemp. Math.* **26**, 1984, pp. 189-206.

[7] Linial N. and London E. and Rabinovich Y. The geometry of graphs and some of its algorithmic applications. *Combinatorica*, **15**, 1995. pp. 215–245. (Preliminary version in : *35th Annual Symposium on Foundations of Computer Science*, 1994, pp. 577–591.)

[8] Matoušek, J. Bi-Lipschitz embeddings into low dimensional Euclidean spaces. *Comment. Math. Univ. Carolinae.* **33**(1), 1992. pp. 51–55.