An Elementary Proof of a Theorem of Johnson and Lindenstrauss

Sanjoy Dasgupta,[1] Anupam Gupta[2]
[1]AT&T Labs Research, Room A277, Florham Park, New Jersey 07932; e-mail: dasgupta@research.att.com
[2]Lucent Bell Labs, Room 2C-355, 600 Mountain Avenue, Murray Hill, New Jersey 07974; e-mail: anupamg@research.bell-labs.com

Received 16 December 2001; accepted 11 July 2002

DOI 10.1002/rsa.10073

ABSTRACT: A result of Johnson and Lindenstrauss [13] shows that a set of n points in high dimensional Euclidean space can be mapped 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 theorem using elementary probabilistic techniques. © 2003 Wiley Periodicals, Inc. Random Struct. Alg., 22: 60–65, 2002

1. INTRODUCTION

A fundamental result of Johnson and Lindenstrauss [13] says that any n point subset of Euclidean space can be embedded in $k = O(\log n/\epsilon^2)$ dimensions without distorting the distances between any pair of points by more than a factor of $(1 \pm \epsilon)$, for any $0 < \epsilon < 1$. In recent work, Noga Alon has shown that this result is essentially tight: His result shows that any set of n points with inter-point distances lying in the range $[1 - \epsilon, 1 + \epsilon]$ requires at least $\Omega(\log n/(\epsilon^2 \log 1/\epsilon))$ dimensions [1, Section 9].

In recent years, the Johnson–Lindenstrauss theorem has found numerous applications that include bi-Lipschitz embeddings of graphs into normed spaces [14], searching for

Correspondence to: A. Gupta

approximate nearest neighbors in high-dimensional Euclidean space [12], learning mixtures of Gaussians [5], and dimension reduction in databases [2].

The original proof of Johnson and Lindenstrauss is probabilistic, showing that projecting the n-point subset onto a random subspace of $O(\log n/\epsilon^2)$ dimensions only changes the interpoint distances by $(1 \pm \epsilon)$ with positive probability. Their proof was subsequently simplified by Frankl and Maehara [7, 8]. The proof given in this note uses elementary probabilistic techniques to obtain the result. Indyk and Motwani [12], Arriaga and Vempala [3], and Achlioptas [2] have also given similar proofs of the theorem using simple randomized algorithms. (A discussion of some of these proofs is given in Section 3.) Many of these randomized algorithms have recently been derandomized by [6, 15].

2. THE JOHNSON-LINDENSTRAUSS THEOREM

The main result of this paper is the following:

Theorem 2.1. *For any $0 < \epsilon < 1$ and any integer n, let k be a positive integer such that*

$$k \geq 4(\epsilon^2/2 - \epsilon^3/3)^{-1} \ln n. \tag{2.1}$$

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

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

Furthermore, this map can be found in randomized polynomial time.

The original paper of Johnson and Lindenstrauss [13] proved a version of this result with the lower bound on k being $O(\log n)$. In their paper, Frankl and Maehara [7] showed that $k = \lceil 9(\epsilon^2 - 2\epsilon^3/3)^{-1} \ln n \rceil + 1$ dimensions are sufficient; the papers of Indyk and Motwani [12] and Achlioptas [2] give essentially the same bounds for k as we do. (For a discussion on these proofs, the reader is pointed to Section 3.)

Our proof of the theorem follows a fairly standard line of reasoning which has been used before for this problem (e.g., in [7]): It shows that the squared length of a random vector is sharply concentrated around its mean when the vector is projected onto a random k-dimensional subspace. Specifically, with probability $O(1/n^2)$, its (scaled) length is not distorted by more than $(1 \pm \epsilon)$. The theorem then follows from a union bound.

Hence the aim is to estimate the length of a unit vector in $\mathbf{R}^d$ 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 Gaussian $N(0, 1)$ random variables, and let $Y = \frac{1}{\|X\|}(X_1, \ldots, X_d)$. It is easy 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 $Z \in \mathbf{R}^k$ be the projection of Y onto its first k coordinates, and let $L = \|Z\|^2$. Clearly the expected squared length of Z

is $\mu = \mathbf{E}[L] = k/d$. The following lemma shows that L is also fairly tightly concentrated around μ.

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

a. If $\beta < 1$, then

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

b. If $\beta > 1$, then

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

Before we prove this lemma, let us see how it implies Theorem 2.1.

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

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

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

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

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

Now set the map $f(v_i) = (\sqrt{d/k})v_i'$. By the above calculations, for some fixed pair i, j, the chance that 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 points suffers a large distortion is at most $\binom{n}{2} \times 2/n^2 = 1 - 1/n$. Hence f has the desired properties with probability at least $1/n$. Repeating this projection $O(n)$ times can boost the success probability to the desired constant, giving us the claimed randomized polynomial time algorithm. ∎

To finish off, let us prove Lemma 2.2. The proof uses by now standard techniques used for proving large deviation bounds on sums of random variables [4, 9].

Proof of Lemma 2.2(a). We use the easily-proved fact that if $X \sim N(0, 1)$, then $E[e^{sX^2}] = 1/\sqrt{1 - 2s}$, for $-\infty < s < \frac{1}{2}$. We now prove that

$$\Pr[d(X_1^2 + \cdots + X_k^2) \le k\beta(X_1^2 + \cdots + X_d^2)] \le \beta^{k/2}\left(1 + \frac{k(1-\beta)}{d-k}\right)^{(d-k)/2}. \quad (2.2)$$

Note that this is just another way of stating Lemma 2.2(a). However, this can be shown by the following algebraic manipulations:

$$\begin{aligned}
&\Pr[d(X_1^2 + \cdots + X_k^2) \le k\beta(X_1^2 + \cdots + X_d^2)] \\
&\quad = \Pr[k\beta(X_1^2 + \cdots + X_d^2) - d(X_1^2 + \cdots + X_k^2) \ge 0] \\
&\quad = \Pr[\exp\{t(k\beta(X_1^2 + \cdots + X_d^2) - d(X_1^2 + \cdots + X_k^2))\} \ge 1] \qquad (\text{for } t > 0) \\
&\quad \le \mathbf{E}[\exp\{t(k\beta(X_1^2 + \cdots + X_d^2) - d(X_1^2 + \cdots + X_k^2))\}] \qquad (\text{by Markov's inequality}) \\
&\quad = \mathbf{E}[\exp\{tk\beta X^2\}]^{(d-k)}\mathbf{E}[\exp\{t(k\beta - d)X^2\}]^k \qquad (\text{where } X \sim N(0,1)) \\
&\quad = (1 - 2tk\beta)^{-(d-k)/2}(1 - 2t(k\beta - d))^{-k/2}.
\end{aligned}$$

We will refer to this last expression as $g(t)$. 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 $t \ge 0$), 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 (2.2). ∎

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

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

for $0 < t < 1/2(d - k\beta)$. But this is minimized at $-t_0$, where t_0 is 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

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

■

3. DISCUSSION

The reader may find it interesting to compare the results of this paper with the alternate proofs given by Indyk and Motwani in [12], and by Achlioptas in [2].

The algorithm in former paper does not choose a random k-dimensional subspace *per se;* it instead picks k independent random vectors $\{U_i\}_{i=1}^k$ from the d-dimensional normal distribution (with the unit covariance matrix), and sets the i-th coordinate of the map $f(x)$ to be $\frac{1}{\sqrt{d}}\langle U_i, x\rangle$. The proof follows by formalizing the intuition that these random vectors are almost orthogonal to each other, and hence this mapping is almost the same as projecting onto a random k-dimensional subspace.

The statement in [12] analogous to our Lemma 2.2 is somewhat weaker in the sense that the lower bound for k contains some lower order terms, as a result of which one has to assume a lower bound for k larger by an additive factor of roughly $O(\log \log n)$. However, their algorithm is substantially simpler, since it just has to populate all the entries of a $k \times d$ matrix A by independent $N(0, 1)$ random variables, whereupon the images of $x \in V$ are given by $f(x)=\frac{1}{\sqrt{d}}(Ax)$.

The latter paper [2] takes this idea even further and shows that, instead of using Gaussians, one can pick the entries of A to be uniformly and independently drawn from $\{1, -1\}$. With a tighter analysis than that of [12], this paper gives the same bound for k as Theorem 2.1.

REFERENCES

[1] N. Alon, Problems and results in extremal combinatorics, Part I, unpublished manuscript.

[2] D. Achlioptas, Database friendly random projections, Proc 20th ACM Symp Principles of Database Systems, Santa Barbara, CA, 2001, 274–281.

[3] R. I. Arriaga and S. Vempala, An algorithmic theory of learning: Robust concepts and random projection, Proc 40th Annu IEEE Symp Foundations of Computer Science, New York, NY, 1999, pp. 616–623.

[4] H. Chernoff, A measure of asymptotic efficiency for tests of a hypothesis based on the sum of observations, Ann Math Stat 23 (1952), 493–507.

[5] S. Dasgupta, Learning mixtures of Gaussians, Proc 40th Annu IEEE Symp Foundations of Computer Science, New York, NY, 1999, pp. 634–644.

[6] L. Engebretsen, P. Indyk, and R. O'Donnell, Derandomized dimensionality reduction with applications, Proc 13th Annu ACM SIAM Symp Discrete Algorithms, San Francisco, CA, 2002, pp. 705–712.

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

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

[9] W. Hoeffding, Probability inequalities for sums of bounded random variables, J Am Stat Assoc 58 (1963), 13–30.

[10] P. Indyk and R. Motwani, Approximate nearest neighbors: Towards removing the curse of dimensionality, Proc 30th Annu ACM Symp Theory of Computing, Dallas, TX, 1998, pp. 604–613.

[11] W. B. Johnson and J. Lindenstrauss, Extensions of Lipschitz maps into a Hilbert space, Contemp Math 26 (1984), 189–206.

[12] N. Linial, F. London, and Y. Rabinovich, The geometry of graphs and some of its algorithmic applications, Combinatorica 15(2) (1995), 215–245 (preliminary version in 35th Annu Symp Foundations of Computer Science, 1994, pp. 577–591).

[13] D. Sivakumar, Algorithmic derandomization using complexity theory, Proc 34th Annu ACM Symp Theory of Computing, Montréal, Canada, 2002, pp. 619–626.