ACADEMIC
PRESS

Available at
www.ComputerScienceWeb.com
POWERED BY SCIENCE @ DIRECT•

Journal of Computer and System Sciences 66 (2003) 671–687

JOURNAL OF COMPUTER AND SYSTEM SCIENCES

http://www.elsevier.com/locate/jcss

Database-friendly random projections: Johnson-Lindenstrauss with binary coins

Dimitris Achlioptas

Microsoft Research, One Microsoft Way, Redmond, WA 98052, USA

Received 28 August 2001; revised 19 July 2002

Abstract

A classic result of Johnson and Lindenstrauss asserts that any set of n points in d-dimensional Euclidean space can be embedded into k-dimensional Euclidean space—where k is logarithmic in n and independent of d—so that all pairwise distances are maintained within an arbitrarily small factor. All known constructions of such embeddings involve projecting the n points onto a spherically random k-dimensional hyperplane through the origin. We give two constructions of such embeddings with the property that all elements of the projection matrix belong in $\{-1, 0, +1\}$. Such constructions are particularly well suited for database environments, as the computation of the embedding reduces to evaluating a single aggregate over k random partitions of the attributes.

1. Introduction

Consider projecting the points of your favorite sculpture first onto the plane and then onto a single line. The result amply demonstrates the power of dimensionality.

In general, given a high-dimensional pointset it is natural to ask if it could be embedded into a lower dimensional space without suffering great distortion. In this paper, we consider this question for finite sets of points in Euclidean space. It will be convenient to think of n points in $\mathbb{R}^d$ as an $n \times d$ matrix A, each point represented as a row (vector).

Given such a matrix representation, one of the most commonly used embeddings is the one suggested by the singular value decomposition of A. That is, in order to embed the n points into $\mathbb{R}^k$ we project them onto the k-dimensional space spanned by the singular vectors corresponding to the k largest singular values of A. If one rewrites the result of this projection as a (rank k) $n \times d$ matrix A_k, we are guaranteed that any other k-dimensional pointset (represented as an $n \times d$

E-mail address: optas@microsoft.com.

0022-0000/03/$ - see front matter © 2003 Elsevier Science (USA). All rights reserved.
doi:10.1016/S0022-0000(03)00025-4

matrix D) satisfies

$$|A - A_k|_F \leqslant |A - D|_F,$$

where, for any matrix Q, $|Q|_F^2 = \sum Q_{i,j}^2$. To interpret this result observe that if moving a point by z takes energy proportional to z^2, A_k represents the k-dimensional configuration reachable from A by expending least energy.

In fact, A_k is an optimal rank k approximation of A under many matrix norms. In particular, it is well-known that for any rank k matrix D and for *any* rotationally invariant norm

$$|A - A_k| \leqslant |A - D|.$$

At the same time, this optimality implies no guarantees regarding *local* properties of the resulting embedding. For example, it is not hard to devise examples where the new distance between a pair of points is arbitrarily smaller than their original distance. For a number of problems where dimensionality reduction is clearly desirable, the absence of such local guarantees can make it hard to exploit embeddings algorithmically.

In a seminal paper, Linial et al. [12] were the first to consider algorithmic applications of embeddings that respect local properties. By now, embeddings of this type have become an important tool in algorithmic design. A real gem in this area has been the following result of Johnson and Lindenstrauss [9].

Lemma 1.1 (Johnson and Lindenstrauss [9]). *Given $\varepsilon > 0$ and an integer n, let k be a positive integer such that $k \geqslant k_0 = O(\varepsilon^{-2} \log n)$. For every set P of n points in $\mathbb{R}^d$ there exists $f : \mathbb{R}^d \to \mathbb{R}^k$ such that for all $u, v \in P$*

$$(1 - \varepsilon)||u - v||^2 \leqslant ||f(u) - f(v)||^2 \leqslant (1 + \varepsilon)||u - v||^2.$$

We will refer to embeddings providing a guarantee akin to that of Lemma 1.1 as JL-embeddings. In the last few years, such embeddings have been used in solving a variety of problems. The idea is as follows. By providing a low-dimensional representation of the data, JL-embeddings speed up certain algorithms dramatically, in particular algorithms whose run-time depends exponentially in the dimension of the working space. (For a number of practical problems the best-known algorithms indeed have such behavior.) At the same time, the provided guarantee regarding pairwise distances often allows one to establish that the solution found by working in the low-dimensional space is a good approximation to the solution in the original space. We give a few examples below.

Papadimitriou et al. [13], proved that embedding the points of A in a low-dimensional space can significantly speed up the computation of a low-rank approximation to A, without significantly affecting its quality. In [8], Indyk and Motwani showed that JL-embeddings are useful in solving the ε-approximate nearest-neighbor problem, where (after some preprocessing of the pointset P) one is to answer queries of the following type: "Given an arbitrary point x, find a point $y \in P$, such that for every point $z \in P$, $||x - z|| \geqslant (1 - \varepsilon)||x - y||$." In a different vein, Schulman [14] used JL-embeddings as part of an approximation algorithm for the version of clustering where we seek to minimize the sum of the squares of intracluster distances. Recently, Indyk [7] showed that

JL-embeddings can also be used in the context of "data-stream" computation, where one has limited memory and is allowed only a single pass over the data (stream).

1.1. Our contribution

Over the years, the probabilistic method has allowed for the original proof of Johnson and Lindenstrauss to be greatly simplified and sharpened, while at the same time giving conceptually simple randomized algorithms for constructing the embedding [5,6,8]. Roughly speaking, all such algorithms project the input points onto a spherically random hyperplane through the origin. While this is conceptually simple, in practical terms it amounts to multiplying the input matrix A with a dense matrix of real numbers. This can be a non-trivial task in many practical computational environments. At the same time, investigating the role of spherical symmetry in the choice of hyperplane is mathematically interesting in itself.

Our main result, below, asserts that one can replace projections onto spherically random hyperplanes with much simpler and faster operations. In particular, these operations can be implemented efficiently in a database environment using standard SQL primitives. Somewhat surprisingly, we prove that this comes without *any* sacrifice in the quality of the embedding. In fact, we will see that for every fixed value of d we get a slightly better bound than all current methods. We state our result below as Theorem 1.1. Similarly to Lemma 1.1, the parameter ε controls the desired accuracy in distance preservation, while now β controls the projection's probability of success.

Theorem 1.1. *Let P be an arbitrary set of n points in $\mathbb{R}^d$, represented as an $n \times d$ matrix A. Given $\varepsilon, \beta > 0$ let*

$$k_0 = \frac{4+2\beta}{\varepsilon^2/2 - \varepsilon^3/3} \log n.$$

For integer $k \geqslant k_0$, let R be a $d \times k$ random matrix with $R(i,j) = r_{ij}$, where $\{r_{ij}\}$ are independent random variables from either one of the following two probability distributions:

$$r_{ij} = \begin{cases} +1 & \text{with probability } 1/2, \\ -1 & \text{with probability } 1/2, \end{cases} \tag{1}$$

$$r_{ij} = \sqrt{3} \times \begin{cases} +1 & \text{with probability } 1/6, \\ 0 & \text{with probability } 2/3, \\ -1 & \text{with probability } 1/6. \end{cases} \tag{2}$$

Let

$$E = \frac{1}{\sqrt{k}} AR$$

and let $f : \mathbb{R}^d \to \mathbb{R}^k$ map the ith row of A to the ith row of E.

With probability at least $1 - n^{-\beta}$, for all $u, v \in P$

$$(1-\varepsilon)||u-v||^2 \leqslant ||f(u) - f(v)||^2 \leqslant (1+\varepsilon)||u-v||^2.$$

We see that to construct a JL-embedding via Theorem 1.1 we only need very simple probability distributions to generate the projection matrix, while the computation of the projection itself reduces to aggregate evaluation, i.e., summations and subtractions (but no multiplications). While it seems hard to imagine a probability distribution much simpler than (1), probability distribution (2) gives an additional threefold speedup as we only need to process a third of all attributes for each of the k coordinates. We note that the construction based on probability distribution (1) was, independently, proposed by Arriaga and Vempala in [2], but without an analysis of its performance.

1.1.1. Projecting onto random lines

Looking a bit more closely into the computation of the embedding we see that each row (vector) of A is projected onto k random vectors whose coordinates $\{r_{ij}\}$ are independent random variables with mean 0 and variance 1. If the $\{r_{ij}\}$ were independent normal random variables with mean 0 and variance 1, it is well-known that each resulting vector would point to uniformly random direction in space. Projections onto such vectors have been considered in a number of settings, including the work of Kleinberg [10] and Kushilevitz et al. [11] on approximate nearest neighbors and of Vempala on learning intersections of halfspaces [16]. More recently, such projections have also been used in learning mixture of Gaussians models, starting with the work of Dasgupta [4] and later with the work of Arora and Kannan [3].

Our proof implies that for any fixed vector α, the behavior of its projection onto a random vector c is mandated by the even moments of the random variable $||\alpha \cdot c||$. In fact, our result follows by showing that for every vector α, under our distributions for $\{r_{ij}\}$, these moments are dominated by the corresponding moments for the case where c is spherically symmetric. As a result, projecting onto vectors whose entries are distributed like the columns of matrix R is computationally simpler and results in projections that are at least as nicely behaved.

1.1.2. Randomization

A naive, perhaps, attempt at constructing JL-embeddings would be to pick k of the original coordinates in d-dimensional space as the new coordinates. Naturally, as two points can be very far apart while only differing along a single dimension, this approach is doomed. Yet, if we knew that for every pair of points all coordinates contributed "roughly equally" to the corresponding pairwise, distance, this naive scheme would make perfect sense.

With this in mind, it is very natural to try and remove pathologies like those mentioned above by first applying a random *rotation* to the original pointset in $\mathbb{R}^d$. Observe now that picking, say, the first k coordinates after applying a random rotation is exactly the same as projecting onto a spherically random k-dimensional hyperplane! Thus, we see that randomization in JL-projections only serves as insurance against axis-alignment, analogous to the application of a random permutation before running Quicksort.

1.1.3. Derandomization

Theorem 1.1 allows one to use significantly fewer random bits than all previous methods for constructing JL-embeddings. Indeed, since the appearance of an earlier version of this work [1], the construction based on probability distribution (1) has been used by Sivakumar [15] to give a very simple derandomization of the Johnson–Lindenstrauss lemma.

2. Previous work

As we will see, in all methods for producing JL-embeddings, including ours, the heart of the matter is showing that for any vector, the squared length of its projection is sharply concentrated around its expected value. The original proof of Johnson and Lindenstrauss [9] uses quite heavy geometric approximation machinery to yield such a concentration bound. That proof was greatly simplified and sharpened by Frankl and Meahara [6] who explicitly considered a projection onto k random orthonormal vectors (as opposed to viewing such vectors as the basis of a random hyperplane), yielding the following result.

Theorem 2.1 (Frankl and Meahara [6]). *For any* $\varepsilon \in (0, 1/2)$, *any sufficiently large set* $P \in \mathbb{R}^d$, *and* $k \geqslant k_0 = \lceil 9(\varepsilon^2 - 2\varepsilon^3/3)^{-1} \log |P| \rceil + 1$, *there exists a map* $f : P \to \mathbb{R}^k$ *such that for all* $u, v \in P$,

$$(1-\varepsilon)||u-v||^2 \leqslant ||f(u) - f(v)||^2 \leqslant (1+\varepsilon)||u-v||^2.$$

The next great simplification of the proof of Lemma 1.1 was given, independently, by Indyk and Motwani [8] and Dasgupta and Gupta [5], the latter also giving a slight sharpening of the bound for k_0. By combining the analysis of [5] with the viewpoint of [8] it is in fact not hard to show that Theorem 1.1 holds if for all $i,j, r_{ij} \overset{\mathrm{D}}{=} N(0,1)$ (this was also observed in [2]). Below we state our rendition of how each of these two latest simplifications [8,5] were achieved, as they prepare the ground for our own work. Let us write $X \overset{\mathrm{D}}{=} y$ to denote that X is distributed as Y and recall that $N(0,1)$ denotes the standard Normal distribution with mean 0 and variance 1.

Indyk and Motwani [8]: Assume that we try to implement the scheme of Frankl and Maehara [6] but we are lazy about enforcing either normality (unit length) or orthogonality among our k vectors. Instead, we just pick k independent, spherically symmetric random vectors, by taking as the coordinates of each vector $ki.i.d.$ $N(0, 1/d)$ random variables (so that the expected length of each vector is 1).

An immediate gain of this approach is that now, for any fixed vector α, the length of its projection onto each of our vectors is also a normal random variable. This is due to a powerful and deep fact, namely the 2-stability of the Gaussian distribution: for any real numbers $\alpha_1, \alpha_2, \ldots, \alpha_d$, if $\{Z_i\}_{i=1}^d$ is a family of independent normal random variables and $X = \sum_{i=1}^d \alpha_i Z_i$, then $X \overset{\mathrm{D}}{=} cN(0,1)$, where $c = (\alpha_1^2 + \cdots + \alpha_d^2)^{1/2}$. As a result, if we take these k projection lengths to be the coordinates of the embedded vector in $\mathbb{R}^k$, then the squared length of the embedded vector follows the chi-square distribution for which strong concentration bounds are readily available.

Remarkably, very little is lost due to this laziness. Although, we did not explicitly enforce either orthogonality, or normality, the resulting k vectors, with high probability, will come very close to having both of these properties. In particular, the length of each of the k vectors is sharply concentrated (around 1) as the sum of d independent random variables. Moreover, since the k vectors point in uniformly random directions in $\mathbb{R}^d$, as d grows they rapidly get closer to being orthogonal.

Dasgupta and Gupta [5]: Here we will exploit spherical symmetry without appealing directly to the 2-stability of the Gaussian distribution. Instead observe that, by symmetry, the projection of any unit vector α on a random hyperplane through the origin is distributed exactly like the projection of a random point from the surface of the d-dimensional sphere onto a fixed subspace of dimension k. Such a projection can be studied readily since each coordinate is a scaled normal random variable. With a somewhat tighter analysis than [8], this approach gave the best known bound, namely $k \geqslant k_0 = (4+2\beta)(\varepsilon^2/2 - \varepsilon^3/3)^{-1} \log n$, which is exactly the same as the bound in Theorem 1.1.

3. Some intuition

Our contribution begins with the realization that spherical symmetry, while making life extremely comfortable, is not essential. What is essential is concentration. So, at least in principle, we are free to consider other candidate distributions for the $\{r_{ij}\}$, if perhaps at the expense of comfort.

As we saw earlier, each column of our projection matrix R will yield one coordinate of the projection in $\mathbb{R}^k$. Thus, the squared length of the projection is merely the sum of the squares of these coordinates. Therefore, the projection can be thought of as follows: each column of R acts as an independent estimator of the original vector's length (its estimate being the inner product with it); to reach consensus we take the sum of the k estimators. Seen from this angle, requiring the k vectors to be orthonormal has the pleasant statistical overtone of "maximizing mutual information" (since all estimators have equal weight and are orthogonal). Nonetheless, even if we only require that each column simply gives an unbiased, bounded variance estimator, the Central Limit Theorem implies that if we take sufficiently many columns, we can get an arbitrarily good estimate of the original length. Naturally, the number of columns needed depends on the variance of the estimators.

From the above we see that the key issue is the concentration of the projection of an arbitrary fixed vector α onto a single random vector. The main technical difficulty that results from giving up spherical symmetry is that this concentration can depend on α. Our technical contribution lies in determining probability distributions for the $\{r_{ij}\}$ under which, for all vectors, this concentration is at least as good as in the spherically symmetric case. In fact, it will turn out that for every *fixed* value of d, we can get a (minuscule) improvement in concentration. Thus, for every fixed d, we can actually get a *strictly better* bound for k than by taking spherically random vectors. The reader might be wondering "how can it be that perfect spherical symmetry does not buy us anything (and is in fact slightly worse for each fixed d)?". The following intuitive argument hints at why giving up spherical symmetry is (at least) not catastrophic.

Say, for example, that $\{r_{ij}\} \in \{-1, +1\}$ so that the projection length of certain vectors is more variable than that of others, and assume that an adversary is trying to pick a worst-case such vector w, i.e., one whose projection length is "most variable." Our problem can be rephrased as "How much are we empowering an adversary by committing to picking our column vectors among lattice points rather than arbitrary points in $\mathbb{R}^d$?". As we will see, and this is the heart of our proof, the worst-case vectors are $\frac{1}{\sqrt{d}}(\pm 1, \ldots, \pm 1)$. So, in some sense, the worst-case vectors are "typical", unlike, say, $(1, 0, \ldots, 0)$. From this it is a small leap to believe that the adversary would not fare much worse by giving us a spherically random vector. But in that case, by symmetry, our commitment to lattice points is irrelevant!

To get a more precise answer it seems like one has to delve into the proof. In particular, both for the spherically random case and for our distributions, the bound on k is mandated by the probability of *over*estimating the projected length. Thus, the "bad events" amount to the spanning vectors being too "well-aligned" with α. Now, in the spherically symmetric setting it is possible to have alignment that is arbitrarily close to perfect, albeit with correspondingly smaller probability. In our case, if we do not have perfect alignment then we are guaranteed a certain, bounded amount of misalignment. It is precisely this tradeoff between the probability and the extent of alignment that drives the proof.

Consider, for example, the case when $d = 2$ with $r_{ij} \in \{-1, +1\}$. As we said above, the worst-case vector is $w = (1/\sqrt{2})(1, 1)$. So, with probability $1/2$ we have perfect alignment (when our random vector is $\pm w$) and with probability $1/2$ we have orthogonality. On the other hand, for the spherically symmetric case, we have to consider the integral over all points on the plane, weighted by their probability under the two-dimensional Gaussian distribution. It is a rather instructive exercise to visually explore this tradeoff by plotting the corresponding functions; moreover, it might offer the interested reader some intuition for the general case.

4. Preliminaries and the spherically symmetric case

4.1. Preliminaries

Let $x \cdot y$ denote the inner product of vectors x, y. To simplify notation in the calculations we will work with a matrix R scaled by $1/\sqrt{d}$. As a result, to get E we need to scale $A \times R$ by $\sqrt{d/k}$ rather than $1/\sqrt{k}$. So, R will be a random $d \times k$ matrix with $R(i,j) = r_{ij}/\sqrt{d}$, where the $\{r_{ij}\}$ are distributed as in Theorem 1.1. Therefore, if c_j denotes the jth column of R, then $\{c_j\}_{j=1}^{k}$ is a family of k i.i.d. random unit vectors in $\mathbb{R}^d$ and for all $\alpha \in \mathbb{R}^d$, $f(\alpha) = \sqrt{d/k}(\alpha \cdot c_1, \ldots, \alpha \cdot c_d)$. Naturally, such scaling can be postponed until after the matrix multiplication (projection) has been performed, so that we maintain the advantage of only having $\{-1, 0, +1\}$ in the projection matrix.

Let us first compute $\mathbf{E}(\|f(\alpha)\|^2)$ for an arbitrary vector $\alpha \in \mathbb{R}^d$. For $j = 1, \ldots, k$ let

$$Q_j(\alpha) = \alpha \cdot c_j,$$

where sometimes we will omit the dependence of $Q_j(\alpha)$ on α and refer simply to Q_j.

Then

$$\mathbf{E}(Q_j) = \mathbf{E}\left(\frac{1}{\sqrt{d}} \sum_{i=1}^{d} \alpha_i r_{ij}\right) = \frac{1}{\sqrt{d}} \sum_{i=1}^{d} \alpha_i \mathbf{E}(r_{ij}) = 0 \tag{3}$$

and

$$\begin{aligned}\mathbf{E}(Q_j^2) &= \mathbf{E}\left(\left(\frac{1}{\sqrt{d}}\sum_{i=1}^{d}\alpha_i r_{ij}\right)^2\right)\\ &= \frac{1}{d}\mathbf{E}\left(\sum_{i=1}^{d}(\alpha_i r_{ij})^2 + \sum_{l=1}^{d}\sum_{m=1}^{d} 2\alpha_l\alpha_m r_{lj} r_{mj}\right)\\ &= \frac{1}{d}\sum_{i=1}^{d}\alpha_i^2\mathbf{E}(r_{ij}^2) + \frac{1}{d}\sum_{l=1}^{d}\sum_{m=1}^{d} 2\alpha_l\alpha_m\mathbf{E}(r_{lj})\mathbf{E}(r_{mj})\\ &= \frac{1}{d}\times \|\alpha\|^2. \end{aligned} \tag{4}$$

Note that to get (3) and (4) we only used that the $\{r_{ij}\}$ are independent with zero mean and unit variance. From (4) we see that

$$\mathbf{E}(\|f(\alpha)\|^2) = \mathbf{E}((\|\sqrt{d/k}(\alpha\cdot c_1,\dots,\alpha\cdot c_d)\|)^2) = \frac{d}{k}\sum_{j=1}^{k}\mathbf{E}(Q_j^2) = \|\alpha\|^2.$$

That is for *any* independent family of $\{r_{ij}\}$ with $\mathbf{E}(r_{ij}) = 0$ and $\mathrm{Var}(r_{ij}) = 1$ we get an unbiased estimator, i.e., $\mathbf{E}(\|f(\alpha)\|^2) = \|\alpha\|^2$.

In order to have a JL-embedding we need that for each of the $\binom{n}{2}$ pairs $u, v \in P$, the squared norm of the vector $u - v$, is maintained within a factor of $1 \pm \varepsilon$. Therefore, if for some family $\{r_{ij}\}$ as above we can prove that for some $\beta > 0$ and any fixed vector $\alpha \in \mathbb{R}^d$,

$$\Pr[(1-\varepsilon)\|\alpha\|^2 \leqslant \|f(\alpha)\|^2 \leqslant (1+\varepsilon)\|\alpha\|^2] \geqslant 1 - \frac{2}{n^{2+\beta}}, \tag{5}$$

then the probability of not getting a JL-embedding is bounded by $\binom{n}{2} \times 2/n^{2+\beta} < 1/n^{\beta}$.

From the above discussion we see that our entire task has been reduced to determining a zero mean, unit variance distribution for the $\{r_{ij}\}$ such that (5) holds for *any* fixed vector α. In fact, since for any fixed projection matrix, $\|f(\alpha)\|^2$ is proportional to $\|\alpha\|^2$, it suffices to prove that (5) holds for arbitrary *unit* vectors. Moreover, since $\mathbf{E}(\|f(\alpha)\|^2) = \|\alpha\|^2$, inequality (5) merely asserts that the random variable $\|f(\alpha)\|^2$ is concentrated around its expectation.

4.2. The spherically symmetric case

As a warmup we first work out the spherically random case below. Our results follows from exactly the same proof after replacing Lemma 4.1 below with a corresponding lemma for our choices of $\{r_{ij}\}$.

Getting a concentration inequality for $\|f(\alpha)\|^2$ when $r_{ij} \overset{\mathrm{D}}{=} N(0,1)$ is straightforward. Due to the 2-stability of the normal distribution, for *every* unit vector α, we have $\|f(\alpha)\|^2 \overset{\mathrm{D}}{=} \chi^2(k)/k$, where $\chi^2(k)$ denotes the chi-square distribution with k degrees of freedom. The fact that we get the same distribution for every vector α corresponds to the obvious intuition that "all vectors are the same" with respect to projection onto a spherically random vector. Standard tail-bounds for the chi-square distribution readily yield the following.

Lemma 4.1. *Let* $r_{ij} \overset{\mathrm{D}}{=} N(0,1)$ *for all* i,j. *Then, for any* $\varepsilon > 0$ *and any unit-vector* $\alpha \in \mathbb{R}^d$,

$$\Pr[\|f(\alpha)\|^2 > 1+\varepsilon] < \exp\left(-\frac{k}{2}(\varepsilon^2/2 - \varepsilon^3/3)\right),$$

$$\Pr[\|f(\alpha)\|^2 < 1-\varepsilon] < \exp\left(-\frac{k}{2}(\varepsilon^2/2 - \varepsilon^3/3)\right).$$

Thus, to get a JL-embedding we need only require

$$2 \times \exp\left(-\frac{k}{2}(\varepsilon^2/2 - \varepsilon^3/3)\right) \leqslant \frac{2}{n^{2+\beta}},$$

which holds for

$$k \geqslant \frac{4+2\beta}{\varepsilon^2/2 - \varepsilon^3/3} \log n.$$

Let us note that the bound on the upper tail of $\|f(\alpha)\|^2$ above is *tight* (up to lower order terms). As a result, as long as the union bound is used, one cannot hope for a better bound on k while using spherically random vectors.

To prove our result we will use the exact same approach, arguing that for every unit vector $\alpha \in \mathbb{R}^d$, the random variable $\|f(\alpha)\|^2$ is sharply concentrated around its expectation. In the next section we state a lemma analogous to Lemma 4.1 above and show how it follows from bounds on certain moments of Q_1^2. We then prove those bounds in Section 6.

5. Tail bounds

To simplify notation let us define for an arbitrary vector α,

$$S = S(\alpha) = \sum_{j=1}^{k} (\alpha \cdot c_j)^2 = \sum_{j=1}^{k} Q_j^2(\alpha),$$

where c_j is the jth column of R, so that $\|f(\alpha)\|^2 = S \times d/k$.

Lemma 5.1. *Let* r_{ij} *have any one of the two distributions in Theorem* 1.1. *Then, for any* $\varepsilon > 0$ *and any unit vector* $\alpha \in \mathbb{R}^d$,

$$\Pr[S(\alpha) > (1+\varepsilon)k/d] < \exp\left(-\frac{k}{2}(\varepsilon^2/2 - \varepsilon^3/3)\right),$$

$$\Pr[S(\alpha) < (1-\varepsilon)k/d] < \exp\left(-\frac{k}{2}(\varepsilon^2/2 - \varepsilon^3/3)\right).$$

In proving Lemma 5.1 we will generally omit the dependence of probabilities on α, making it explicit only when it affects our calculations. We will use the standard technique of applying Markov's inequality to the moment generating function of S, thus reducing the proof of the lemma to bounding certain moments of Q_1. In particular, we will need the following lemma which will be proved in Section 6.

Lemma 5.2. *For all* $h \in [0, d/2)$, *all* $d \geqslant 1$ *and all unit vectors* α,

$$\mathbf{E}(\exp(hQ_1(\alpha)^2)) \leqslant \frac{1}{\sqrt{1-2h/d}}, \tag{6}$$

$$\mathbf{E}(Q_1(\alpha)^4) \leqslant \frac{3}{d^2}. \tag{7}$$

Proof of Lemma 5.1. We start with the upper tail. For arbitrary $h > 0$ let us write

$$\begin{aligned}\Pr\left[S > (1+\varepsilon)\frac{k}{d}\right] &= \Pr\left[\exp(hS) > \exp\left(h(1+\varepsilon)\frac{k}{d}\right)\right]\\ &< \mathbf{E}(\exp(hS))\exp\left(-h(1+\varepsilon)\frac{k}{d}\right).\end{aligned}$$

Since $\{Q_j\}_{j=1}^k$ are i.i.d. we have

$$\mathbf{E}(\exp(hS)) = \mathbf{E}\left(\prod_{j=1}^{k}\exp(hQ_j^2)\right) \tag{8}$$

$$= \prod_{j=1}^{k}\mathbf{E}(\exp(hQ_j^2)) \tag{9}$$

$$= (\mathbf{E}(\exp(hQ_1^2)))^k, \tag{10}$$

where passing from (8) to (9) uses that the $\{Q_j\}_{j=1}^k$ are independent, while passing from (9) to (10) uses that they are identically distributed. Thus, for any $\varepsilon > 0$

$$\Pr\left[S > (1+\varepsilon)\frac{k}{d}\right] < (\mathbf{E}(\exp(hQ_1^2)))^k \exp\left(-h(1+\varepsilon)\frac{k}{d}\right). \tag{11}$$

Substituting (6) in (11) we get (12). To optimize the bound we set the derivative in (12) with respect to h to 0. This gives $h = \frac{d}{2}\frac{\varepsilon}{1+\varepsilon} < \frac{d}{2}$. Substituting this value of h we get (13) and series expansion yields (14).

$$\Pr\left[S > (1+\varepsilon)\frac{k}{d}\right] < \left(\frac{1}{\sqrt{1-2h/d}}\right)^k \exp\left(-h(1+\varepsilon)\frac{k}{d}\right) \tag{12}$$

$$= ((1+\varepsilon)\exp(-\varepsilon))^{k/2} \tag{13}$$

$$< \exp\left(-\frac{k}{2}(\varepsilon^2/2 - \varepsilon^3/3)\right). \tag{14}$$

Similarly, but now considering $\exp(-hS)$ for arbitrary $h>0$, we get that for any $\varepsilon>0$

$$\Pr\left[S<(1-\varepsilon)\frac{k}{d}\right]<(\mathbf{E}(\exp(-hQ_1^2)))^k\exp\left(h(1-\varepsilon)\frac{k}{d}\right). \tag{15}$$

Rather than bounding $\mathbf{E}(\exp(-hQ_1^2))$ directly, let us expand $\exp(-hQ_1^2)$ to get

$$\begin{aligned}\Pr\left[S<(1-\varepsilon)\frac{k}{d}\right]&<\left(\mathbf{E}\left(1-hQ_1^2+\frac{(-hQ_1^2)^2}{2!}\right)\right)^k\exp\left(h(1-\varepsilon)\frac{k}{d}\right)\\&=\left(1-\frac{h}{d}+\frac{h^2}{2}\mathbf{E}(Q_1^4)\right)^k\exp\left(h(1-\varepsilon)\frac{k}{d}\right),\end{aligned} \tag{16}$$

where $\mathbf{E}(Q_1^2)$ was given by (4).

Now, substituting (7) in (16) we get (17). This time taking $h=\frac{d}{2}\frac{\varepsilon}{1+\varepsilon}$ is not optimal but is still "good enough", giving (18). Again, series expansion yields (19).

$$\Pr\left[S<(1-\varepsilon)\frac{k}{d}\right]\leqslant\left(1-\frac{h}{d}+\frac{3}{2}\left(\frac{h}{d}\right)^2\right)^k\exp\left(h(1-\varepsilon)\frac{k}{d}\right) \tag{17}$$

$$=\left(1-\frac{\varepsilon}{2(1+\varepsilon)}+\frac{3\varepsilon^2}{8(1+\varepsilon)^2}\right)^k\exp\left(\frac{\varepsilon(1-\varepsilon)k}{2(1+\varepsilon)}\right) \tag{18}$$

$$<\exp\left(-\frac{k}{2}(\varepsilon^2/2-\varepsilon^3/3)\right).\qquad\square \tag{19}$$

6. Moment bounds

To simplify notation in this section we will drop the subscript and refer to Q_1 as Q. It should be clear that the distribution of Q depends on α, i.e., $Q=Q(\alpha)$. This is precisely what we give up by not projecting onto spherically symmetric vectors. Our strategy for giving bounds on the moments of Q will be to determine a "worst-case" unit vector w and bound the moments of $Q(w)$. We claim the following.

Lemma 6.1. *Let*

$$w=\frac{1}{\sqrt{d}}(1,\ldots,1).$$

For every unit vector $\alpha\in\mathbb{R}^d$, and for all $k=0,1,\ldots$

$$\mathbf{E}(Q(\alpha)^{2k})\leqslant\mathbf{E}(Q(w)^{2k}). \tag{20}$$

Moreover, we will prove that the even moments of $Q(w)$ are dominated by the corresponding moments from the spherically symmetric case. That is,

Lemma 6.2. *Let $T \overset{\mathrm{D}}{=} N(0, 1/d)$. For all $d \geqslant 1$ and all $k = 0, 1, \ldots$*

$$\mathbf{E}(Q(w)^{2k}) \leqslant \mathbf{E}(T^{2k}). \tag{21}$$

Using Lemmata 6.1 and 6.2 we can prove Lemma 5.2 as follows.

Proof of Lemma 5.2. To prove (7) we observe that for any unit vector α, by (20) and (21),

$$\mathbf{E}(Q(\alpha)^4) \leqslant \mathbf{E}(Q(w)^4) \leqslant \mathbf{E}(T^4),$$

while

$$\mathbf{E}(T^4) = \int_{-\infty}^{+\infty} \frac{1}{\sqrt{2\pi}} \exp(-\lambda^2/2) \left(\frac{\lambda^4}{d^2}\right) d\lambda = \frac{3}{d^2}.$$

To prove (6) we first observe that for any real-valued random variable U and for all h such that $\mathbf{E}(\exp(hU^2))$ is bounded, the Monotone Convergence Theorem (MCT) allows us to swap the expectation with the sum and get

$$\mathbf{E}(\exp(hU^2)) = \mathbf{E}\left(\sum_{k=0}^{\infty} \frac{(hU^2)^k}{k!}\right) = \sum_{k=0}^{\infty} \frac{h^k}{k!} \mathbf{E}(U^{2k}).$$

So, below, we proceed as follows. Taking $h \in [0, d/2)$ makes the integral in (22) converge, giving us (23). Thus, for such h, we can apply the MCT to get (24). Now, applying (20) and (21)–(24) gives (25). Applying the MCT once more gives (26).

$$\mathbf{E}(\exp(hT^2)) = \int_{-\infty}^{+\infty} \frac{1}{\sqrt{2\pi}} \exp(-\lambda^2/2) \exp\left(h\frac{\lambda^2}{d}\right) d\lambda \tag{22}$$

$$= \frac{1}{\sqrt{1 - 2h/d}} \tag{23}$$

$$= \sum_{k=0}^{\infty} \frac{h^k}{k!} \mathbf{E}(T^{2k}) \tag{24}$$

$$\geqslant \sum_{k=0}^{\infty} \frac{h^k}{k!} \mathbf{E}(Q(\alpha)^{2k}) \tag{25}$$

$$= \mathbf{E}(\exp(hQ(\alpha)^2)). \tag{26}$$

Thus, $\mathbf{E}(\exp(hQ^2)) \leqslant 1/\sqrt{1 - 2h/d}$ for $h \in [0, d/2)$, as desired. □

Before proving Lemma 6.1 we will need to prove the following lemma.

Lemma 6.3. *Let r_1, r_2 be i.i.d. random variables having one of the two probability distributions given by Eqs.* (1) *and* (2) *in Theorem* 1.1.

For any $a, b \in \mathbb{R}$ let $c = \sqrt{(a^2+b^2)/2}$. Then for any $M \in \mathbb{R}$ and all $k = 0, 1, \ldots$

$$\mathbf{E}((M + ar_1 + br_2)^{2k}) \leqslant \mathbf{E}((M + cr_1 + cr_2)^{2k}).$$

Proof. We first consider the case where $r_i \in \{-1, +1\}$.

If $a^2 = b^2$ then $a = c$ and the lemma holds with equality. Otherwise, observe that

$$\mathbf{E}((M + cr_1 + cr_2)^{2k}) - \mathbf{E}((M + ar_1 + br_2)^{2k}) = \frac{S_k}{4},$$

where

$$S_k = (M + 2c)^{2k} + 2M^{2k} + (M - 2c)^{2k} - (M + a + b)^{2k} \\ - (M + a - b)^{2k} - (M - a + b)^{2k} - (M - a - b)^{2k}.$$

We will show that $S_k \geqslant 0$ for all $k \geqslant 0$.

Since $a^2 \neq b^2$ we can use the binomial theorem to expand every term other than $2M^{2k}$ in S_k and get

$$S_k = 2M^{2k} + \sum_{i=0}^{2k} \binom{2k}{i} M^{2k-i} D_i,$$

where

$$D_i = (2c)^i + (-2c)^i - (a + b)^i - (a - b)^i - (-a + b)^i - (-a - b)^i.$$

Observe now that for odd i, $D_i = 0$. Moreover, we claim that $D_{2j} \geqslant 0$ for all $j \geqslant 1$. To see this claim observe that $(2a^2 + 2b^2) = (a + b)^2 + (a - b)^2$ and that for all $j \geqslant 1$ and $x, y \geqslant 0$, $(x + y)^j \geqslant x^j + y^j$. Thus, $(2c)^{2j} = (2a^2 + 2b^2)^j = [(a + b)^2 + (a - b)^2]^j \geqslant (a + b)^{2j} + (a - b)^{2j}$ implying

$$S_k = 2M^{2k} + \sum_{j=0}^{k} \binom{2k}{2j} M^{2(k-j)} D_{2j} = \sum_{j=1}^{k} \binom{2k}{2j} M^{2(k-j)} D_{2j} \geqslant 0.$$

The proof for the case where $r_i \in \{-\sqrt{3}, 0, +\sqrt{3}\}$ is just a more cumbersome version of the proof above, so we omit it. That proof, though, brings forward an interesting point. If one tries to take $r_i = 0$ with probability greater than 2/3, while maintaining that r_i has a range of size 3 and variance 1, the lemma fails. In other words, 2/3 is tight in terms of how much probability mass we can put to $r_i = 0$ and still have the current lemma hold. □

Proof of Lemma 6.1. Recall that for any vector α, $Q(\alpha) = Q_1(\alpha) = \alpha \cdot c_1$ where

$$c_1 = \frac{1}{\sqrt{d}}(r_{11}, \ldots, r_{d1}).$$

If $\alpha = (\alpha_1, \ldots, \alpha_d)$ is such that $\alpha_i^2 = \alpha_j^2$ for all i, j, then by symmetry, $Q(\alpha)$ and $Q(w)$ are identically distributed and the lemma holds trivially. Otherwise, we can assume without loss of generality, that $\alpha_1^2 \neq \alpha_2^2$ and consider the "more balanced" unit vector $\theta = (c, c, \alpha_3, \dots, \alpha_d)$, where $c = \sqrt{(\alpha_1^2 + \alpha_2^2)/2}$. We will prove that

$$\mathbf{E}(Q(\alpha)^{2k}) \leqslant \mathbf{E}(Q(\theta)^{2k}). \tag{27}$$

Applying this argument repeatedly yields the lemma, as θ eventually becomes w.

To prove (27), below we first express $\mathbf{E}(Q(\alpha)^{2k})$ as a sum of averages over r_{11}, r_{21} and then apply Lemma 6.3 to get that each term (average) in the sum, is bounded by the corresponding average for vector θ. More precisely,

$$\begin{aligned}\mathbf{E}(Q(\alpha)^{2k}) &= \frac{1}{d^k} \sum_M \mathbf{E}((M + \alpha_1 r_{11} + \alpha_2 r_{21})^{2k}) \Pr\left[\sum_{i=3}^{d} \alpha_i r_{i1} = \frac{M}{\sqrt{d}}\right] \\ &\leqslant \frac{1}{d^k} \sum_M \mathbf{E}((M + c r_{11} + c r_{21})^{2k}) \Pr\left[\sum_{i=3}^{d} \alpha_i r_{i1} = \frac{M}{\sqrt{d}}\right] \\ &= \mathbf{E}(Q(\theta)^{2k}). \qquad \square\end{aligned}$$

Proof of Lemma 6.2. Recall that $T \overset{\mathrm{D}}{=} N(0, 1/d)$. We will first express T as the scaled sum of d independent standard Normal random variables. This will allow for a direct comparison of the terms in each of the two expectations.

Specifically, let $\{T_i\}_{i=1}^d$ be a family of i.i.d. standard Normal random variables. Then $\sum_{i=1}^d T_i$ is a Normal random variable with variance d. Therefore,

$$T \overset{\mathrm{D}}{=} \frac{1}{d} \sum_{i=1}^{d} T_i.$$

Recall also that $Q(w) = Q_1(w) = w \cdot c_1$ where

$$c_1 = \frac{1}{\sqrt{d}} (r_{11}, \dots, r_{d1}).$$

To simplify notation let us write $r_{i1} = Y_i$ and let us also drop the dependence of Q on w. Thus,

$$Q = \frac{1}{d} \sum_{i=1}^{d} Y_i,$$

where $\{Y_i\}_{i=1}^d$ are i.i.d. r.v. having one of the two distributions in Theorem 1.1.

We are now ready to compare $\mathbf{E}(Q^{2k})$ with $\mathbf{E}(T^{2k})$. We first observe that for every $k = 0, 1, \dots$

$$\mathbf{E}(T^{2k}) = \frac{1}{d^{2k}} \sum_{i_1=1}^{d} \cdots \sum_{i_{2k}=1}^{d} \mathbf{E}(T_{i_1} \cdots T_{i_{2k}})$$

and

$$\mathbf{E}(Q^{2k}) = \frac{1}{d^{2k}} \sum_{i_1=1}^{d} \cdots \sum_{i_{2k}=1}^{d} \mathbf{E}(Y_{i_1} \cdots Y_{i_{2k}}).$$

To prove the lemma we will show that for every value assignment to the indices $i_1, \ldots, i_{2k}$,

$$\mathbf{E}(Y_{i_1} \cdots Y_{i_{2k}}) \leqslant \mathbf{E}(T_{i_1} \cdots T_{i_{2k}}). \tag{28}$$

Let $V = \langle v_1, v_2, \ldots, v_{2k} \rangle$ be the value assignment considered. For $i \in \{1, \ldots, d\}$, let $c_V(i)$ be the number of times that i appears in V. Observe that if for some i, $c_V(i)$ is odd then both expectations appearing in (28) are 0, since both $\{Y_i\}_{i=1}^d$ and $\{T_i\}_{i=1}^d$ are independent families and $\mathbf{E}(Y_i) = \mathbf{E}(T_i) = 0$ for all i. Thus, we can assume that there exists a set $\{j_1, j_2, \ldots, j_p\}$ of indices and corresponding values $\ell_1, \ell_2, \ldots, \ell_p$ such that

$$\mathbf{E}(Y_{i_1} \cdots Y_{i_{2k}}) = \mathbf{E}(Y_{j_1}^{2\ell_1} Y_{j_2}^{2\ell_2} \cdots Y_{j_p}^{2\ell_p})$$

and

$$\mathbf{E}(T_{i_1} \cdots T_{i_{2k}}) = \mathbf{E}(T_{j_1}^{2\ell_1} T_{j_2}^{2\ell_2} \cdots T_{j_p}^{2\ell_p}).$$

Note now that since the indices $j_1, j_2, \ldots, j_p$ are distinct, $\{Y_{j_t}\}_{t=1}^p$ and $\{T_{j_t}\}_{t=1}^p$ are families of i.i.d. r.v. Therefore,

$$\mathbf{E}(Y_{i_1} \cdots Y_{i_{2k}}) = \mathbf{E}(Y_{j_1}^{2\ell_1}) \times \cdots \times \mathbf{E}(Y_{j_p}^{2\ell_p})$$

and

$$\mathbf{E}(T_{i_1} \cdots T_{i_{2k}}) = \mathbf{E}(T_{j_1}^{2\ell_1}) \times \cdots \times \mathbf{E}(T_{j_p}^{2\ell_p}).$$

So, without loss of generality, in order to prove (28) it suffices to prove that for every $\ell = 0, 1, \ldots$

$$\mathbf{E}(Y_1^{2\ell}) \leqslant \mathbf{E}(T_1^{2\ell}). \tag{29}$$

This, though, is completely trivial. First recall the well-known fact that the (2ℓ)th moment of $N(0,1)$ is $(2\ell - 1)!! = (2\ell)!/(\ell! 2^\ell) \geqslant 1$. Now:

- *If* $Y_1 \in \{-1, +1\}$ then $\mathbf{E}(Y_1^{2\ell}) = 1$, for all $\ell \geqslant 0$.
- If $Y_1 \in \{-\sqrt{3}, 0, +\sqrt{3}\}$ then $\mathbf{E}(Y_1^{2\ell}) = 3^{\ell-1} \leqslant (2\ell)!/(\ell! 2^\ell)$, where the last inequality follows by an easy induction.

It is worth pointing out that, along with Lemma 6.3, these are the only two points were we used any properties of the distributions for the r_{ij} (here called Y_i) other than them having zero mean and unit variance. □

Finally, we note that

- Since $\mathbf{E}(Y_1^{2\ell}) < \mathbf{E}(T_1^{2\ell})$ for certain l, we see that for each fixed d, both inequalities in Lemma 5.2 are actually strict, yielding slightly better tails bounds for S and a correspondingly better bound for k_0.

- By using Jensen's inequality one can get a direct bound for $\mathbf{E}(Q^{2k})$ when $Y_i \in \{-1, +1\}$, i.e., without comparing it to $\mathbf{E}(T^{2k})$. That simplifies the proof for that case and shows that, in fact, taking $Y_i \in \{-1, +1\}$ is the minimizer of $\mathbf{E}(\exp(hQ^2))$ for all h.

7. Discussion

7.1. Database-friendliness

As we mentioned earlier, all previously known constructions of JL-embeddings required the multiplication of the input matrix with a dense, random matrix. Unfortunately, such general matrix multiplication can be very inefficient (and cumbersome) to carry out in a relational database.

Our constructions, on the other hand, replace the inner product operations of matrix multiplication with view selection and aggregation (addition). Using, say, distribution (2) of Theorem 1.1 the k new coordinates are generated by independently performing the following random experiment k times: throw away $2/3$ of the original attributes at random; partition the remaining attributes randomly into two parts; for each part, produce a new attribute equal to the sum of all its attributes; take the difference of the two sum attributes.

7.2. Further work

It is well-known that $\Omega(\varepsilon^{-2} \log n / \log(1/\varepsilon))$ dimensions are necessary for embedding arbitrary sets of n points with distortion $1 \pm \varepsilon$. At the same time, all currently known embeddings amount to (random) projections requiring $O(\varepsilon^{-2} \log n)$ dimensions. As we saw, the current analysis of such projections is tight, except for using the union bound to bound the total probability of the $\binom{n}{2}$ potential bad events. Exploring the possibility of savings on this point appears interesting, especially since in practice we often can make *some* assumptions about the input points. For example, what happens if the points are guaranteed to already lie (approximately) in some low-dimensional space?

Finally, we note that with effort (and using a different proof) one can increase the probability of 0 in distribution (2) slightly above $2/3$. Yet, it seems hard to get a significant increase without incurring a penalty in the dimensionality. Exploring the suggested tradeoff might also be interesting, at least for practice purposes.

Acknowledgments

I am grateful to Marek Biskup for his help with the proof of Lemma 6.2 and to Jeong Han Kim for suggesting the approach of Eq. (16). Many thanks also to Paul Bradley, Aris Gionis, Anna Karlin, Elias Koutsoupias and Piotr Indyk for comments on earlier versions of the paper and useful discussions.

References

[1] D. Achlioptas, Database-friendly random projections, 20th Annual Symposium on Principles of Database Systems, Santa Barbara, CA, 2001, pp. 274–281.

[2] R.I. Arriaga, S. Vempala, An algorithmic theory of learning: robust concepts and random projection, 40th Annual Symposium on Foundations of Computer Science, New York, NY, 1999, IEEE Computer Society Press, Los Alamitos, CA, 1999, pp. 616–623.

[3] S. Arora, R. Kannan, Learning mixtures of arbitrary Gaussians, 33rd Annual ACM Symposium on Theory of Computing, Creete, Greece, ACM, New York, 2001, pp. 247–257.

[4] S. Dasgupta, Learning mixtures of Gaussians, 40th Annual Symposium on Foundations of Computer Science, New York, NY, 1999, IEEE Computer Society Press, Los Alamitos, CA, 1999, pp. 634–644.

[5] S. Dasgupta, A. Gupta, An elementary proof of the Johnson–Lindenstrauss lemma. Technical Report 99-006, UC Berkeley, March 1999.

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

[7] P. Indyk, Stable distributions, pseudorandom generators, embeddings and data stream computation, 41st Annual Symposium on Foundations of Computer Science, Redondo Beach, CA, 2000, IEEE Computer Society Press, Los Alamitos, CA, 2000, pp. 189–197.

[8] P. Indyk, R. Motwani, Approximate nearest neighbors: towards removing the curse of dimensionality, 30th Annual ACM Symposium on Theory of Computing, Dallas, TX, ACM, New York, 1998, pp. 604–613.

[9] W.B. Johnson, J. Lindenstrauss, Extensions of Lipschitz mappings into a Hilbert space, Conference in modern analysis and probability, New Haven, CI, 1982, Amer. Math. Soc., Providence, RI, 1984, pp. 189–206.

[10] J. Kleinberg, Two algorithms for nearest-neighbor search in high dimensions, 29th Annual ACM Symposium on Theory of Computing, El Paso, TX, 1997, ACM, New York, 1997, pp. 599–608.

[11] E. Kushilevitz, R. Ostrovsky, Y. Rabani, Efficient search for approximate nearest neighbor in high dimensional spaces, SIAM J. Comput. 30 (2) (2000) 457–474.

[12] N. Linial, E. London, Y. Rabinovich, The geometry of graphs and some of its algorithmic applications, Combinatorica 15 (2) (1995) 215–245.

[13] C.H. Papadimitriou, P. Raghavan, H. Tamaki, S. Vempala, Latent semantic indexing: a probabilistic analysis, 17th Annual Symposium on Principles of Database Systems, Seattle, WA, 1998, pp. 159–168.

[14] L.J. Schulman, Clustering for edge-cost minimization, 32nd Annual ACM Symposium on Theory of Computing, Portland, OR, 2000, ACM, New York, 2000, pp. 547–555.

[15] D. Sivakumar, Algorithmic derandomization via complexity theory, 34th Annual ACM Symposium on Theory of Computing, Montreal, QC, 2002, ACM, New York, 2002, pp. 619–626.

[16] S. Vempala, A random sampling based algorithm for learning the intersection of half-spaces, 38th Annual Symposium on Foundations of Computer Science, Miami, FL, 1997, IEEE Computer Society Press, Los Alamitos, CA, 1997, pp. 508–513.