Discrete Comput Geom (2009) 41: 273–283
DOI 10.1007/s00454-008-9068-8

Lower Bounds for Local Versions of Dimension Reductions

Gideon Schechtman · Adi Shraibman

Received: 6 September 2007 / Revised: 19 December 2007 /
Published online: 28 February 2008

Abstract We consider the problem of embedding vectors from an arbitrary Euclidean space into a low-dimensional Euclidean space while preserving, up to a small distortion, a subset of the distances. In particular, preserving only the distance of each vector to a small number of its nearest neighbors. We show that even when the subset of distances we wish to preserve is very small, the problem does not become easier than when one is required to preserve *all* the distances.

Keywords Johnson–Lindenstrauss lemma · Dimension reduction · Nearest neighbors

1 Introduction

We consider the problem of embedding vectors from an arbitrary Euclidean space into a low-dimensional Euclidean space while preserving a subset of the distances. Formally, given a graph $G=(V,E)$ on $V=[n]=\{1,2,\dots,n\}$, $\varepsilon>0$ and vectors $x_1,\dots,x_n\in\mathbb{R}^n$, we ask: What is the minimal k such that there are vectors $y_1,\dots,y_n\in\mathbb{R}^k$ satisfying $(1-\varepsilon)\|x_i-x_j\|\le\|y_i-y_j\|\le(1+\varepsilon)\|x_i-x_j\|$ for every $(i,j)\in E$? Concerning the nonedges, we consider two paradigms. The first asks for some lower bound on $\|y_i-y_j\|$ for $(i,j)\notin E$, and the second imposes no restriction on these distances.

The motivation for considering this problem comes from a particularly interesting instance of it, when the underlined graph is the graph induced by the m nearest

Supported in part by the Israel Science Foundation.

G. Schechtman (✉) · A. Shraibman
Department of Mathematics, Weizmann Institute of Science, Rehovot, Israel
e-mail: gideon.schechtman@weizmann.ac.il

A. Shraibman
e-mail: adi.shraibman@weizmann.ac.il

neighbors of $x_1, \dots, x_n$. That is, $(i, j) \in E$ if and only if x_j is among the m nearest neighbors to x_i or x_i is among the m nearest neighbors to x_j. We shall call this graph the m-nearest neighbors graph.

A well-known result of Johnson and Lindenstrauss [4] asserts that for the complete graph K_n, k can be taken to be $O(\frac{\log n}{\varepsilon^2})$. An example due to Alon [2] shows that, for K_n, this is basically the best possible; in this example k must be $\Omega(\frac{\log n}{\varepsilon^2 \log 1/\varepsilon})$. We remark in passing that it is unknown whether the Johnson–Lindenstrauss estimate can be improved in terms of the dependence on ε or whether one can find an example which necessarily satisfies $k = \Omega(\frac{\log n}{\varepsilon^2})$.

One might hope that if one only wants to preserve the, say, three nearest neighbors of each vector, one could do with a much smaller k. The main purpose of this note is to show that this is not the case.

One can interpret the phrase “preserve the m-nearest neighbors” in two ways. First, one may want to preserve, up to a factor of $1+\varepsilon$, the distances $\|x_i - x_j\|$ whenever x_j is among the m-nearest neighbors to x_i or x_i is among the m-nearest neighbors to x_j and one does not care what happen to the other distances. Second, one may want to additionally ensure that, for each i and j, y_j is among the first m neighbors of y_i if and only if x_j is among the first m neighbors of x_i. That is, the map sending the x_i's to the y_i's preserves—in addition to the approximate distances—the graph of the m-nearest neighbors.

In the second, more interesting case, we show a simple example (with $m = 2$) in which necessarily $k = \Omega(\frac{\log n}{\varepsilon^2 \log 1/\varepsilon})$; see Theorem 8. In the first case (not caring what happens to nonclosest neighbors) we build, in Theorem 9, a somewhat more elaborate example (with $m = 3$) such that for each sufficiently small $\varepsilon > 0$, k still needs to be $\Omega(\log n)$. In contrast to the example discussed above, here we do not get good dependence on ε.

The proof of Theorem 8 depends on a theorem evaluating the rank of positive semidefinite matrices which are perturbations of the identity matrix. This result, which is a variation of a result of Alon [2], is discussed in Sect. 2. We find it to be of independent interest and we discuss some more applications a-la-Alon [2] in Sect. 6.

Our attention to the problems discussed in this paper was prompted by Abraham, Bartal and Neiman [1]. Some problems raised there are solved here (see in particular the last sentence in [1]).

2 The Rank of Some Positive Semidefinite Matrices

As mentioned in the introduction, the best lower bound on dimension reduction in the general case is due to Alon [2]. Alon's proof relies on a general lower bound for the rank of a real matrix in which the absolute value of all off-diagonal entries is significantly smaller than the value of diagonal entries (Theorem 1.1 in [2]).

We need a somewhat stronger version of Theorem 1.1 from [2] for positive semidefinite matrices. The point is that, in this case, we only need an upper bound on the value of off-diagonal entries, instead of an upper bound on their absolute value.

Theorem 1 *Let A be an $n \times n$ symmetric positive semi-definite matrix such that $a_{ii} \geq 1/2$ for all i and $a_{ij} \leq \varepsilon$ for all $i \neq j$, where $\frac{1}{n^{1/3}} \leq \varepsilon \leq 1/4$. Then*

$$\operatorname{rank}(A) \geq \Omega\left(\frac{\log n}{\varepsilon^2 \log 1/\varepsilon}\right).$$

Note that the range of ε here is a bit smaller than that in Alon's theorem (the lower bound is $n^{-1/3}$ rather than $n^{-1/2}$). Alon [2] gives many applications of his Theorem 1.1, many of which involve semi-definite matrices. For these applications one can use Theorem 1 and get a somewhat stronger assertion. We give two examples in Sect. 6.

The proof of Theorem 1 is similar to the proof of Theorem 1.1 in [2], but there are some non-trivial modifications. The most substantial change is to replace Lemma 2.2 from [2] with Lemma 4 below.

As in [2], it is convenient to first prove the following theorem.

Theorem 2 *Let A be an $n \times n$ symmetric positive-definite matrix such that $a_{ii} = 1$ for all i and $a_{ij} \leq \varepsilon$ for all $i \neq j$, where $\frac{1}{n^{1/3}} \leq \varepsilon \leq 1/2$. Then*

$$\operatorname{rank}(A) \geq \Omega\left(\frac{\log n}{\varepsilon^2 \log 1/\varepsilon}\right).$$

Remark 3 We note that Theorem 2 is equivalent to giving a lower bound on the dimension of a spherical code with a given number of elements (equivalently, an upper bound on the maximal size of spherical codes of a given dimension). A k-dimensional spherical code of angular distance θ and size n is an n element subset of the $k-1$-dimensional sphere such that the angle between any two points in the set is at least θ. Equivalently, the inner product between any two points is at most $\cos(\theta)$. Good asymptotic bounds for the maximal size of a spherical code with fixed distance θ are known; see, e.g., [3]. The advantage of our bound (on the dimension k) is that it holds for every ε and n, and that the proof is elementary. We discuss this further in Sect. 6.1.

We need the following two lemmas.

Lemma 4 *Let A be an $n \times n$ symmetric positive semidefinite matrix such that $a_{ij} \leq \varepsilon$ for every $1 \leq i < j \leq n$, and $a_{ii} = 1$ for all $i = 1, \ldots, n$. Then*

$$\operatorname{rank}(A) \geq \frac{n}{2(\varepsilon^{-1} + (n-1)\varepsilon^2)}.$$

In particular, if $\epsilon \leq \frac{1}{n^{1/3}}$, then $\operatorname{rank}(A) \geq n^{2/3}/4$.

Proof Recall that given two symmetric positive semidefinite matrices of the same dimensions, (b_{ij}) and (c_{ij}), the matrix $(b_{ij}c_{ij})$ is also symmetric positive semidefinite. Since both A and (a_{ij}^2) are symmetric positive semidefinite, it follows that,

$$\sum_{i,j=1}^{n} a_{ij} a_{ij}^2 \geq 0.$$

Put $I=\{(i,j); a_{i,j}<-\varepsilon\}$, it follows that

$$-\sum_{(i,j)\in I} a_{i,j}a_{i,j}^2 \le \sum_{(i,j)\in I^c} a_{i,j}a_{i,j}^2 \le \sum_{\{(i,j);\ a_{i,j}>0\}} a_{i,j}a_{i,j}^2 \le n+\varepsilon\cdot \sum_{\{(i,j);\ i\neq j,\ a_{i,j}>0\}} a_{i,j}^2,$$

which implies that

$$\sum_{(i,j)\in I} a_{i,j}^2 \le n\varepsilon^{-1} + \sum_{\{(i,j);\ i\neq j,\ a_{i,j}>0\}} a_{i,j}^2,$$

concluding that $\sum_{i,j=1}^n a_{i,j}^2 \le 2n(\varepsilon^{-1}+(n-1)\varepsilon^2)$.

The claim now follows by using the following bound. For every real matrix X,

$$\operatorname{rank}(X) \ge \Big(\sum x_{ii}\Big)^2 \Big/ \Big(\sum x_{ij}^2\Big). \qquad \square$$

Lemma 5 ([2]) *Let $B=(b_{i,j})$ be an $n\times n$ matrix of rank d, and let $P(x)$ be an arbitrary polynomial of degree k. Then the rank of the $n\times n$ matrix $P(b_{i,j})$ is at most $\binom{k+d}{k}$. Moreover, if $P(x)=x^k$, then the rank of $(P(b_{i,j}))$ is at most $\binom{k+d-1}{k}$.*

Proof of Theorem 2 The proof is quite similar to the proof of Theorem 2.1 in [2]. We repeat it here for completeness.

Denote $d=\operatorname{rank}(A)$. If $\varepsilon\le n^{-\delta}$ for any fixed $0<\delta<1/3$ the result follows from Lemma 4, thus we may assume that $\varepsilon\ge n^{-\delta}$ for some fixed small $\delta>0$. Let $k=\lfloor\frac{\log n}{3\log 1/\varepsilon}\rfloor$ if this number is odd, and $k=\lfloor\frac{\log n}{3\log 1/\varepsilon}\rfloor-1$ otherwise, so that k is always an odd positive integer. The assumption that $\varepsilon\ge n^{-\delta}$ ensures that k is larger than any fixed integer we wish. Set also $n'=\lfloor\varepsilon^{-3k}\rfloor$ and note that $n'\le n$ and $\varepsilon^k\le\frac{1}{(n')^{1/3}}$.

Denote by B the $n'\times n'$ principle minor of the matrix $(a_{i,j}^k)$. B is symmetric positive semidefinite, its diagonal elements are all 1, and off-diagonal elements are all smaller than $\varepsilon^k\le\frac{1}{(n')^{1/3}}$. By Lemma 5 the rank of B is at most $\binom{d+k}{k}\le(\frac{e(k+d)}{k})^k$. On the other hand, by Lemma 4 the rank of B is at least $(n')^{2/3}/4$. Thus

$$\left(\frac{e(k+d)}{k}\right)^k \ge (n')^{2/3}/4 = \frac{1}{4}\left(\left\lfloor\frac{1}{\varepsilon^{3k}}\right\rfloor\right)^{2/3} \ge \frac{1}{8\varepsilon^{2k}}.$$

The result now follows from simple algebraic manipulations. □

Proof of Theorem 1 Let $d=\operatorname{rank}(A)$, then $a_{ij}=\langle x_i,x_j\rangle$ for some set $x_1,\dots,x_n\in\mathbb{R}^d$ and all $1\le i\le j\le n$. Take $y_i=\frac{x_i}{\|x_i\|_2}$ for $i=1,\dots,n$, and consider the matrix $B=(b_{i,j})$ where $b_{i,j}=\langle y_i,y_j\rangle$. B is positive semidefinite, its diagonal elements all equal to 1, its off diagonal elements are at most 2ε and $\operatorname{rank}(B)\le\operatorname{rank}(A)$. The result thus follows from Theorem 2. □

Remark 6 As is clear from the proof, the numbers 1/2 and 1/4 in the statement of Theorem 1 can be replaced by any two numbers $a>b>0$. This will affect only the unspecified value of the constant hiding behind the Ω notation in the conclusion of the theorem.

3 Preserving the Nearest Neighbors Graph

We begin with a very simple result showing that one cannot preserve the closest neighbor (i.e., $m = 1$) and the 1-nearest neighbor(s) graph unless $k = \Omega(\log n)$. This happens even if we allow large distortion and relax somewhat the requirement of preserving the 1-nearest neighbors graph.

Proposition 7 *For each n there are $n + 1$ points $0 = x_0, x_1, \dots, x_n$ in $\mathbb{R}^n$ such that all the distances $\|x_i - x_j\|$ are different and such that if $y_0, y_1, \dots, y_n$ are vectors in $\mathbb{R}^k$ with*

$$a^{-1}\|x_i - x_{i^*}\| \le \|y_i - y_{i^*}\| \le b\|x_i - x_{i^*}\|$$

for all i, where x_{i^*} is the closest to x_i, and

$$\|y_i - y_{i^*}\| \le c\|y_i - y_j\|$$

for some $c > 0$, all i and all $j \ne i, i^*$. Then $k \ge \frac{\log n}{\log(1+2abc^*)}$ where $c^* = \max(c, 1)$.

Proof Let A be any symmetric $n \times n$ real matrix with $1 \ge a_{i,i} \ge 1 - \delta$ and $|a_{i,j}| \le \delta$ for all $1 \le i < j \le n$. If $\delta > 0$ is small enough this matrix is positive definite, so there are $x_1, \dots, x_n$ in $\mathbb{R}^n$ with $\langle x_i, x_j\rangle = a_{i,j}$. If δ is small enough $\|x_i - x_j\| > 1$ for all $i \ne j; i, j = 1, \dots, n$. In particular 0 is closer to x_i, than any other x_j. It is easy to arrange that in addition all the distances $\|x_i - x_j\|$, $i \ne j, i, j = 0, \dots, n$ are different (with $x_0 = 0$).

Let $\{y_i\}_{i=0}^n$ be as in the statement of Proposition 7 and assume as we may that $y_0 = 0$, then $y_{i^*} = 0$ for all $i = 1, \dots, n$ and thus $\|y_i\| \le b$, $i = 1, \dots, n$. Also, $\|y_i - y_j\| \ge (ac^*)^{-1}$ for all $i \ne j; i, j = 1, \dots, n$.

We get that the n balls of radius $1/2ac^*$ centered at the y_i's are disjoint and that all are contained in a ball of radius $b + \frac{1}{2ac^*}$. Consequently, $n(\frac{1}{2ac^*})^k \le (b + \frac{1}{2ac^*})^k$ or

$$k \ge \frac{\log n}{\log(1 + 2abc^*)}. \qquad \square$$

Note that the result above does not give good dependence on ε even if, as we shall assume below, $1 \le a, b \le 1 + \varepsilon$ and $c = 1$. Indeed, the best lower bound one could expect from this result is $\log n/\log 3$. Moreover, it is easy to see that the example built in the proof above does not give anything better (except maybe for the numerical constant $\log 3$). This follows from the fact that, for some numerical constant $c > 1$, there are c^k unit vectors in $\mathbb{R}^k$ such that the distance between each two of them is between 1.1 and 1.2, say. We shall remedy this in the following result, which is the main result of this section, by considering the first two nearest neighbors.

Theorem 8 *For each n and $\frac{1}{n^{1/3}} \le \varepsilon \le \frac{1}{14}$ there are $n + 1$ points $0 = x_0, x_1, \dots, x_n$ in $\mathbb{R}^n$ such that all the distances $\|x_i - x_j\|$ are different and such that if $y_0, y_1, \dots, y_n$ are vectors in $\mathbb{R}^k$ with*

$$(1 + \varepsilon)^{-1}\|x_i - x_{i^*}\| \le \|y_i - y_{i^*}\| \le (1 + \varepsilon)\|x_i - x_{i^*}\|,$$

and

$$(1+\varepsilon)^{-1}\|x_i - x_{i^{**}}\| \le \|y_i - y_{i^{**}}\| \le (1+\varepsilon)\|x_i - x_{i^{**}}\|$$

for all i, where x_{i^*} (resp. $x_{i^{**}}$) denotes the closest (resp. second closest) vector to x_i, and

$$\|y_i - y_{i^{**}}\| \le \|y_i - y_j\|$$

for all i and all $j \ne i, i^*$. Then $k \ge \Omega(\frac{\log n}{\varepsilon^2 \log(1/\varepsilon)})$.

Proof As in the previous proof, let A be any symmetric $n \times n$ real matrix with $1 \ge a_{i,i} \ge 1-\delta$ and $|a_{i,j}| \le \delta$ for all $1 \le i < j \le n$. If $\delta > 0$ is small enough, this matrix is positive definite, so there are $x_1, \dots, x_n$ in $\mathbb{R}^n$ with $\langle x_i, x_j \rangle = a_{i,j}$. If δ is small enough $\|x_i\| \ge (1+\varepsilon)^{-1}$ and $\|x_i - x_j\| > \sqrt{2}(1+\varepsilon)^{-1}$ for all $i \ne j; i, j = 1, \dots, n$. In particular 0 is closer to x_i than any other x_j. It is easy to arrange that in addition all the distances $\|x_i - x_j\|$, $i \ne j; i, j = 0, \dots, n$ are different (with $x_0 = 0$).

Let $\{y_i\}_{i=0}^n$ be as in the statement of the proposition and assume as we may that $y_0 = 0$, then $y_{i^*} = 0$ for all $i = 1, \dots, n$ and thus $(1+\varepsilon)^{-2} \le \|y_i\| \le 1+\varepsilon$, $i = 1, \dots, n$. Also,

$$\|y_i - y_j\| \ge \|y_i - y_{i^{**}}\| \ge \sqrt{2}(1+\varepsilon)^{-2}$$

for all $i \ne j; i, j = 1, \dots, n$.

Renormalizing the y_i's and looking at their Gram matrix we get an at-most rank k symmetric positive semidefinite matrix with 1's on the diagonal and with off-diagonal entries bounded from above by $O(\varepsilon)$. (Note, however, that the absolute value of some of the entries may be large.) We conclude the proof using Theorem 1. □

4 Preserving Only the Nearest Neighbors

Here we show that even if we only want to preserve the three nearest neighbors, with no requirement on the other distances, we cannot, in general, reduce the dimension below $\Omega(\log n)$. We do not get a good dependence on ε here.

Theorem 9 *There is an $\varepsilon > 0$ and $c > 0$ such that for each n there are $n+1$ points $0 = x_0, x_1, \dots, x_n$ in $\mathbb{R}^n$ such that all the distances $\|x_i - x_j\|$ are different and such that if $y_0, y_1, \dots, y_n$ are vectors in $\mathbb{R}^k$ with*

$$(1+\varepsilon)^{-1}\|x_i - x_{i_j}\| < \|y_i - y_{i_j}\| < (1+\varepsilon)\|x_i - x_{i_j}\|$$

for all i and $j = 1, 2, 3$ where $x_{i_1}, x_{i_2}, x_{i_3}$ denotes the three closest vectors to x_i. Then $k \ge c \log n$.

Proof Assume as we may that $n = h(h+1)/2$ for some h. Let $\{e_i\}_{i=1}^h \cup \{e_{i,j}\}_{1 \le i < j \le h}$ be an orthonormal basis of $\mathbb{R}^n$.

Put $z_0 = 0$, $z_i = e_i$, $i = 1, \dots, h$, $z_{i,j} = \frac{2}{3}e_i + \frac{1}{3}e_j + 2e_{i,j}$, $1 \le i < j \le h$. Note that for all $1 \le i < j \le h$

1. $\|z_i - z_j\| = \sqrt{2}$, $\|z_i\| = 1$.
2. $\|z_{i,j} - z_i\| = \sqrt{\frac{2}{9} + 4}$, $\|z_{i,j} - z_j\| = \sqrt{\frac{8}{9} + 4}$, $\|z_{i,j}\| = \sqrt{\frac{5}{9} + 4}$.
3. $\|z_{i,j} - z_s\| = \sqrt{\frac{5}{9} + 5}$, $s \neq i, j, 0$.
4. $\|z_{i,j} - z_{u,v}\| > \sqrt{8}$, $(u, v) \neq (i, j)$.

It follows in particular that the closest to each z_i among the z_j's and $z_{u,v}$'s is z_0. It also follows that the closest to each $z_{i,j}$ is z_i, the second closest z_0 and the third z_j. If we want all distances to be different we may now perturb the vectors an arbitrarily small perturbation while keeping the structure of the first three closest neighbors. We call the new sequence x_t, $t = i$ for $i = 0, \dots, h$ or $t = i, j$ for $1 \leq i < j \leq h$.

Assume now y_t are in $\mathbb{R}^k$ and satisfy

$$(1+\varepsilon)^{-1}\|x_t - x_s\| < \|y_t - y_s\| < (1+\varepsilon)\|x_t - x_s\|$$

for all t and all s such that x_s is among the first three closest neighbors of x_t. Then, assuming the perturbation is small enough, $\|y_i\| \leq 1 + \varepsilon$ for all $i = 1, \dots, h$ and

$$\|y_j - y_i\| \geq \|y_j - y_{i,j}\| - \|y_i - y_{i,j}\| \geq (1+\varepsilon)^{-1}\sqrt{\frac{8}{9} + 4} - (1+\varepsilon)\sqrt{\frac{2}{9} + 4}$$

for all $1 \leq i < j \leq h$. If ε is small enough, this last quantity is larger than 0.01.

We now proceed as in the proof of Proposition 7: We have h disjoint balls of radius 0.005 contained in a ball of radius 2 (say) around y_0, so $h(0.005)^k < 2^k$ and thus $k > \frac{\log h}{\log 400} > \frac{\log n}{2\log 400}$. □

Yair Bartal informed us that Itai Abraham, Ofer Neiman and himself had a similar idea for an example of an n-point metric space (not clearly a subset of Euclidean space) satisfying the conclusion of Theorem 9.

There are two related problems to which we do not know the answer:

1. What is the behavior of the best dimension k as a function of ε as $\varepsilon \to 0$? Does an estimate like the one we get in Theorem 8 still hold?
2. What happens for large distortion? That is, when ε is allowed to be some large constant? Is there then an upper bound on k which is better than $O(\log n)$? We remark that Theorem 8 in [1] gives such a bound under additional assumptions.

5 Other Graphs

Recall that we consider the problem of embedding vectors $x_1, \dots, x_n \in \mathbb{R}^n$ into a low-dimensional Euclidean space, while preserving a subset of the distances. We identify the subset of pairs (x_i, x_j) whose distances we wish to preserve with the edges of a graph $G = (V, E)$ on $V = [n] = \{1, 2, \dots, n\}$. Denote by $k_\varepsilon(G, \{x_i\})$ the minimal k such that there are vectors $y_1, \dots, y_n \in \mathbb{R}^k$ satisfying $(1-\varepsilon)\|x_i - x_j\| \leq \|y_i - y_j\| \leq (1+\varepsilon)\|x_i - x_j\|$ for every $(i, j) \in E$. Also, let $k^1_\varepsilon(G, \{x_i\})$ be the minimal k such that there are vectors $y_1, \dots, y_n \in \mathbb{R}^k$ satisfying $(1-\varepsilon)\|x_i - x_j\| \leq \|y_i - y_j\| \leq (1+\varepsilon)\|x_i - x_j\|$ for every $(i, j) \in E$, and $\|y_i - y_j\| \geq \max_{\{s;(i,s)\in E\}}\|y_i - y_s\|$, for every i, j such that $(i, j) \notin E$.

In Sect. 3 we exhibited graphs G and $x_0, x_1, \dots, x_n \in \mathbb{R}^n$ such that

$$k_\varepsilon^1\big(G, \{x_i\}\big) \geq \Omega\left(\frac{\log n}{\varepsilon^2 \log(1/\varepsilon)}\right),$$

for every $\frac{1}{n^{1/3}} \leq \varepsilon \leq \frac{1}{14}$. In Sect. 4 we gave an example of graphs G and $x_0, x_1, \dots, x_n \in \mathbb{R}^n$ such that $k_\varepsilon(G, \{x_i\}) \geq \Omega(\log n)$ for small enough ε. All these graphs had very large degree, at least at one vertex, and this was important in our analysis. Note, however, that the graphs we have considered up to this point can naturally be viewed also as directed graphs (where each vertex is connected by an outgoing edge to its m-nearest neighbors), these graphs then have a constant and low out-degree but some vertices may have large in-degree. Here we consider the value of k_ε^1 for d-regular graphs. In Theorem 11 we give an example of a set of vectors $X \subset \mathbb{R}^n$ such that $k_\varepsilon^1(G, X)$ is large for expander graphs.

We use the following Poincaré-type inequality (see, e.g., [5, p. 548]).

Lemma 10 *Let $G = ([n], E)$ be a d-regular graph with second eigenvalue bound λ. Then for every $x_1, \dots, x_n \in \mathbb{R}^n$ it holds that*

$$\frac{d-\lambda}{dn^2} \sum_{(i,j)\in[n]} \|x_i - x_j\| \leq |E|^{-1} \sum_{(i,j)\in E} \|x_i - x_j\|.$$

Theorem 11 *Denote by $e_1, \dots, e_n$ the standard basis in $\mathbb{R}^n$, and fix $0 < \varepsilon < 1/2$. If $G = ([n], E)$ is a d-regular graph, $d \geq 3$, with second eigenvalue bounded by $d/2$, then there is no set of vectors $y_1, \dots, y_n \in \mathbb{R}^k$ satisfying*

$$(1-\varepsilon)\|e_i - e_j\| \leq \|y_i - y_j\|, \tag{1}$$

for every pair $i, j \in [n]$, and

$$(1-\varepsilon)\|e_i - e_j\| \leq \|y_i - y_j\| \leq (1+\varepsilon)\|e_i - e_j\|, \tag{2}$$

for every $(i, j) \in E$, unless $k \geq \Omega(\log n)$.

Proof Let $y_1, \dots, y_n \in \mathbb{R}^k$ satisfy the conditions of the theorem, we show that $k \geq \Omega(\log n)$. It follows from our assumptions and Lemma 10 that

$$\frac{1}{n^2} \sum_{i,j\in[n]} \|y_i - y_j\| \leq \frac{2}{|E|} \sum_{(i,j)\in E} \|y_i - y_j\| \leq (1+\varepsilon)\sqrt{8}.$$

The first inequality is by Lemma 10, and the second uses (2).

Also, there always exists a point y_o such that

$$\frac{1}{n} \sum_{r\in[n]} \|y_o - y_r\| \leq \frac{1}{n^2} \sum_{i,j\in[n]} \|y_i - y_j\| \leq (1+\varepsilon)\sqrt{8}.$$

Assume without loss of generality that $y_o = 0$, then

$$\frac{1}{n}\sum_{i\in[n]}\|y_i\| \le (1+\varepsilon)\sqrt{8}.$$

We get that there are at most $n/\sqrt{2}$ vectors y_i with $\|y_i\| \ge 4(1+\varepsilon)$.

We thus have $(1-\frac{1}{\sqrt{2}})n$ y_i's whose norms are at most 6 and whose mutual distances at least $1/\sqrt{2}$ (by (1)). We can now conclude as in the proof of Proposition 7. □

Remarks

1. We note that the assumption on the second eigenvalue in Theorem 11 is essential. Consider the graph $G = ([n]^d, E)$, where $(u,v) \in E$ if and only if $\|u-v\|_1 = 1$. The set of vertices of G are naturally embedded in $\mathbb{R}^d$, as a subset of the integers grid, in such a manner that the distance between every two adjacent vertices in G is exactly 1, and all other distances are greater than 1. This graph has maximal degree $2d$, but is not quite $2d$ regular. A similar example of a $2d$ regular graph is the following: Consider a cycle C_n of length n. It is naturally embedded in $\mathbb{R}^2$ (on an actual circle), where the Euclidean distances of edges are exactly one. Put $G = (C_n^d, E)$, where $((u_1,\dots,u_d),(v_1,\dots,v_d)) \in E$ iff the two vectors differ exactly in one coordinate i and (u_i, v_i) is an edge in C_n. G is $2d$ regular and it naturally embeds into $\mathbb{R}^{2d}$ where the distance between adjacent vertices is 1 and the other distances are at least $\sqrt{2}$.
2. The assumption in Theorem 11 that the embedding is noncontractible is also unavoidable. This follows from the following upper bound. Given a graph $G = (V,E)$ on n vertices and $\varepsilon > 0$, $e_1,\dots,e_n$ are embedded in R^k with $k = \Omega(\log\chi(G)/\varepsilon^2)$ while preserving the distances on the edges of G up to $(1\pm\varepsilon)$. To see this, consider a coloring of G into $\chi(G)$ colors. First map $\{e_i\}$ to the standard basis $\{\hat{e}_a\}$ in $\mathbb{R}^{\chi(G)}$ by mapping e_i to $\hat{e}_a$, where a is the color of the vertex i. It is not hard to see that this mapping preserves the distances on the edges of G. We then use the Johnson–Lindenstrauss lemma with the set $\{\hat{e}_a\}$ to reduce the dimension to $\Omega(\log\chi(G)/\varepsilon^2)$.

   Since the chromatic number of d-regular graphs is at most $d+1$, the above (canonical) example does not provide a good lower bound when we allow the distances on nonedges to change arbitrarily.
3. Finally we comment that the lower bound in Theorem 11 does not depend on ε. Therefore, when ε is very small there might still be a better upper bound than for the complete graph.

6 More Applications of Theorem 1

Alon gives in [2] several applications of his Theorem 1.1 other than for distortion in low-dimension embedding. For many of these applications Theorem 1 can be used to give slightly more general results (for a more restrictive set of ε's). We give here two examples, and refer the reader to [2] for more details.

6.1 Coding Theory

A *binary code* of length k is a set $C \subset \{\pm 1\}^k$. We say that a code is ε*-good* if the Hamming distance between any two code-words is at least $\frac{1-\varepsilon}{2}k$. A major challenge in coding theory is to determine the cardinality of a largest ε-good code of length k.

A code is called ε*-balanced* if it is ε-good and in addition the Hamming distance between any two code words is at most $\frac{1+\varepsilon}{2}k$. The following upper bound on the size of ε-balanced codes was proved in [2].

Claim 12 [2] *There exists an absolute positive constant α so that for all $\frac{1}{\sqrt{k}} \le \varepsilon \le 1/2$ the cardinality of any ε-balanced code of length k is at most $2^{\alpha\varepsilon^2 \log(1/\varepsilon)k}$.*

Using Alon's proof but applying Theorem 1 and Lemma 4 instead of the corresponding counterparts we get a similar bound for ε-good codes.

Claim 13 *There exists an absolute positive constant α so that for all $\frac{1}{k^{1/3}} \le \varepsilon \le 1/2$ the cardinality of any ε-good code of length k is at most $2^{\alpha\varepsilon^2 \log(1/\varepsilon)k}$.*

Proof Let $C = \{c_1, \dots, c_n\} \subset \{\pm 1\}^k$ be an ε-good code, and denote by B the matrix whose rows are the vectors $c_i/\sqrt{k}$, $i = 1, \dots, n$. The matrix BB^t is symmetric positive semidefinite, its diagonal elements are all 1, and its off-diagonal elements are all smaller than ε. Thus by Theorem 1

$$k \ge \operatorname{rank}(BB^t) \ge \frac{\log n}{\alpha\varepsilon^2 \log 1/\varepsilon},$$

for some positive constant α. The claim easily follows. □

In fact, as noted in Sect. 4, the statements of Theorem 1 and Lemma 4 can be interpreted as the following lower bounds for spherical codes. The lower bounds for binary codes are a special case of this fact, since binary codes can be seen as a special case of spherical codes (by normalizing the sign vectors).

Claim 14 *Fix $0 \le \theta \le 1$, and denote by $M(k,\theta)$ the maximal size of a k-dimensional spherical code with distance θ, then if* $\cos(\theta) \ge 1/k^{1/3}$

$$\frac{1}{k} \log M(k,\theta) \le O\big(\cos(\theta)^2 \log \cos(\theta)\big).$$

6.2 Nearly Independent Random Variables

Let $X_1, \dots, X_n$ be a set of random variables over a sample space S of size m, attaining values in $\{0,1\}$. For every subset $Y \subset [n]$, denote by X_Y the xor of the X_i's, $i \in Y$; i.e. $X_Y = \bigoplus_{i \in Y} X_i$. The family $X_1, \dots, X_n$ is called ε*-biased* if for every $Y \ne \emptyset$,

$$\big|\Pr[X_Y = 0] - \Pr[X_Y = 1]\big| \le \varepsilon.$$

Theorem 15 [2] *Let $\{X_1, \dots, X_n\}$ be an ε-biased set of n random variables over a sample space of size m. If $\varepsilon \geq 2^{-n/2}$, then $m \geq \Omega(\frac{n}{\varepsilon^2 \log 1/\varepsilon})$. If $\varepsilon < 2^{-n/2}$, then $m \geq \Omega(2^n)$.*

The proof uses rank lower bounds, as in Sect. 6.1. Since the underlined matrix is again symmetric positive semidefinite, the same lower bound holds, with essentially the same proof, even if we only assume that for every $Y \neq \emptyset$, $\Pr[X_Y = 0] \leq \Pr[X_Y = 1] + \varepsilon$. That is:

Theorem 16 *Let $\{X_1, \dots, X_n\}$ be a set of n binary random variables over a sample space of size m satisfying for every $Y \neq \emptyset$,*

$$\Pr[X_Y = 0] \leq \Pr[X_Y = 1] + \varepsilon.$$

Then $m \geq \Omega(\frac{n}{\varepsilon^2 \log 1/\varepsilon})$ if $\varepsilon \geq 2^{-n/3}$.

References

1. Abraham, I., Bartal, Y., Neiman, O.: Local embeddings of metric spaces. In: STOC, pp. 631–640. Assoc. Comput. Mach., New York (2007)
2. Alon, N.: Perturbed identity matrices have high rank: Proof and applications. Comb. Probab. Comput. (2008, to appear)
3. Conway, J.H., Sloane, N.J.A.: Sphere Packings, Lattices and Groups, 3rd edn. Grundlehren der Mathematischen Wissenschaften [Fundamental Principles of Mathematical Sciences], vol. 290. Springer, New York (1999). With additional contributions by Bannai, E., Borcherds, R.E., Leech, J., Norton, S.P., Odlyzko, A.M., Parker, R.A., Queen, L., Venkov, B.B.
4. Johnson, W.B., Lindenstrauss, J.: Extensions of Lipschitz mappings into a Hilbert space. In: Conference in modern analysis and probability (New Haven Conn., 1982). Contemp. Math., vol. 26, pp. 189–206. Am. Math. Soc., Providence (1984)
5. Linial, N., Hoory, S., Wigderson, A.: Expander graphs and their applications. Bull. Am. Math. Soc. **43**(4), 439–561 (2006)