On Approximating the Average Distance Between Points

Kfir Barhum   Oded Goldreich*   Adi Shraibman

Faculty of Mathematics and Computer Science
Weizmann Institute of Science, Rehovot, Israel.

April 26, 2007

Abstract

We consider the problem of approximating the average distance between pairs of points in a high-dimensional Euclidean space, and more generally in any metric space. Our aim is providing linear-time approximation algorithms, which in particular beat the obvious quadratic-time algorithm that computes the exact value. We consider two algorithmic approaches:

1. Referring only to Euclidean Spaces, we randomly reduce the high-dimensional problem to a one-dimensional problem. The resulting algorithm runs in *almost-linear time* and lends itself to a "direct" derandomization.
2. An alternative approach consists of selecting a random sample of pairs of points and outputting the average distance between these pairs. It turns out that, for any metric space, it suffices to use a sample of size that is linear in the number of points. A "direct" derandomization of this algorithm is meaningless, but we study possible "indirect" derandomization that are inspired by it. In the special case of Euclidean Spaces, we obtain a derandomization result that is superior to the one obtained by the first approach.

Thus, it seems that the second approach is superior to the first approach.

Keywords: General metric spaces, high-dimensional Euclidean spaces, randomized reductions, derandomization, metric embeddings, edge expansion, approximating average quantities.

*Partially supported by the Israel Science Foundation (grant No. 460/05).

Introduction

As observed by Feige [2], natural objects give rise to functions for which approximating the average value of a function is easier than approximating the average value of a general function with a corresponding domain and range. For example, the average degree of a connected n-vertex graph can be approximated upto some constant factor (i.e., 2) based on $\sqrt{n}$ samples, whereas the average value of a general function from $[n]$ to $[n-1]$ cannot be approximated to within any constant factor based on $o(n)$ samples. Indeed, the discrepancy is due to the restrictions imposed on functions that represent quantities that correspond to the type of object considered (i.e., degrees of a graph).

Goldreich and Ron initiated a general study of approximating average parameters of graphs [3]. In particular, they considered the problems of approximating the average degree of a vertex in a graph as well as approximating the average distance between pairs of vertices. They considered both queries to the quantity of interest (e.g., the degree of a vertex) and natural queries to the corresponding object (e.g., neighborhood queries in a graph). (Barhum [1, Chap. 2] extended their average-degree approximation algorithm to k-uniform hypergraphs.)

In the present paper, we consider the problem of approximating the average distance between points in a (high-dimensional) Euclidean space, and more generally for points in any metric space. Although this study may be viewed as an imitation of [3], the specific context (i.e., geometry rather than graph theory) is different and indeed different techniques are employed.

Our aim is beating the obvious quadratic-time algorithm, which may be used to compute the exact value of the aforementioned average. But, unlike in the graph theoretic setting (cf. [3]), we cannot hope for sublinear-time approximation algorithms (because a single "exceptional" point may dominate the value of the average of all pairwise distances). Thus, we seek (almost) linear-time approximation algorithms. We consider two algorithmic approaches.

1. *Manipulating the object itself.* This algorithmic approach (presented in Section 1) applies only to the case of Euclidean Spaces. The algorithm operates by randomly reducing the high-dimensional problem to a one-dimensional problem. The resulting algorithm runs in *almost-linear time* and lends itself to a "direct" derandomization (which incurs an overhead that is exponential in the original dimension).

2. *Sampling and averaging.* The straightforward approach (presented in Section 2) consists of selecting a random sample of pairs of points and outputting the average distance between these pairs. This algorithm works for any metric space, and it turns out that it suffices to use a sample of size that is linear (rather than quadratic) in the number of points. This algorithm does not offer a meaningful "direct" derandomization (because such a derandomization requires scanning all pairs of points), but it raises the question of whether a fixed set of pairs can be used as a basis for approximation. We study this question in Sections 2.2 and 2.3, demonstrating a gap between the case of general metric and the case of Euclidean metric. In particular, in the latter case, we obtain a deterministic approximation algorithm running (essentially) in linear-time.

Thus, it seems that the second algorithmic approach is superior to the first approach. Furthermore, we believe that Sections 2.2 and 2.3 may be of independent interest. In particular, they yield a simple proof to the fact that the graph metric of most constant-degree graphs cannot be embedded in a Euclidean space without incurring logarithmic distortion.

1 Euclidean Spaces and the Random Projection Algorithm

In this section, we present an almost linear time algorithm for approximating the sum of the distances between points in a high-dimensional Euclidean space; that is, given $P_1, ..., P_n \in \mathsf{R}^d$, the algorithm outputs an approximation of $\sum_{i,j\in[n]} \|P_i - P_j\|$. The algorithm is based on randomly reducing the high-dimensional case to the one-dimensional case, where the problem is easily solvable. Specifically, the d-dimensional algorithm repeatedly selects a uniformly distributed direction, projects all points to the corresponding line, and computes the sum of the corresponding distances (on this line). Each such experiment yields an *expected* value that is a $\rho(d)$ fraction of the sum that we seek, where $\rho(d)$ denotes the expected length of the projection of a uniformly distributed unit vector on a fixed direction. Furthermore, $O(\epsilon^{-2})$ repetitions suffice for obtaining a $1 \pm \epsilon$ factor approximation (with error probability at most 1/3). The process can be derandomized at a cost of slowing down the algorithm by a factor of $\text{poly}(1/\epsilon)^d$.

1.1 The one-dimensional case

Our starting point is the fact that an almost linear-time algorithm that computes the exact value (of the sum of all pairwise distances) is known in the one-dimensional case. This algorithm proceeds by first sorting the input points $p_1, ..., p_n \in \mathsf{R}$ such that $p_1 \le p_2 \le \cdots \le p_n$, then computing $\sum_{j=1}^n |p_1 - p_j|$ (in a straightforward manner), and finally for $i = 1, ..., n-1$ computing in constant-time the value $\sum_{j=1}^n |p_{i+1} - p_j|$ based on $\sum_{j=1}^n |p_i - p_j|$. Specifically, we use the fact that

$$\begin{aligned}
\sum_{j=1}^n |p_{i+1} - p_j| &= \sum_{j=1}^i (p_{i+1} - p_j) + \sum_{j=i+1}^n (p_j - p_{i+1}) \\
&= (i - (n-i)) \cdot (p_{i+1} - p_i) + \sum_{j=1}^i (p_i - p_j) + \sum_{j=i+1}^n (p_j - p_i) \\
&= (2i - n) \cdot (p_{i+1} - p_i) + \sum_{j=1}^n |p_i - p_j| .
\end{aligned}$$

1.2 A simple deterministic approximation for the d-dimensional case

Combining the aforementioned algorithm with the basic inequalities regarding norms (i.e., the relation of Norm2 to Norm1), we immediately obtain a (deterministic) $\sqrt{d}$-factor approximation algorithm for the d-dimensional case. Specifically, consider the points $P_1, ..., P_n \in \mathsf{R}^d$, where $P_i = (p_{i,1}, ..., p_{i,d})$, and let $\|P_i - P_j\|$ denote the Euclidean (i.e., Norm2) distance between P_i and P_j. Then it holds that

$$\frac{1}{\sqrt{d}} \cdot \sum_{i,j\in[n]} \sum_{k=1}^d |p_{i,k} - p_{j,k}| \le \sum_{i,j\in[n]} \|P_i - P_j\| \le \sum_{i,j\in[n]} \sum_{k=1}^d |p_{i,k} - p_{j,k}| . \tag{1}$$

Thus, $\sum_{i,j\in[n]} \|P_i - P_j\|$ can be approximated by $\sum_{i,j\in[n]} \sum_{k=1}^d |p_{i,k} - p_{j,k}|$, which is merely the sum of d one-dimensional problems (i.e., $\sum_{k=1}^d \sum_{i,j\in[n]} |p_{i,k} - p_{j,k}|$).

1.3 The main algorithm

However, we seek a better approximation than the $\sqrt{d}$-approximation just described. Indeed, the main contribution of this section is an almost linear-time (randomized) approximation scheme for the value of $\sum_{i,j\in[n]} \|P_i - P_j\|$. The key conceptual observation is that the rough bounds provided by Eq. (1) reflect (extremely different) worst-case situations, whereas "on the average" there is a tight relation between the Norm2 and the Norm1 values. Recall that while the Norm2 value is invariant of the system of coordinates, Norm1 is defined based on such a system and is very dependent on it. What we are saying is that rather than computing the Norm1 value according to an arbitrary system of coordinates (which leaves some slackness w.r.t the Norm2 value that we seek), we should compute the Norm1 value according to a random system of coordinates (i.e., a system that is selected uniformly at random).

To see what will happen when we use a random system of coordinates (i.e., orthonormal basis of R^d), we need some notation. Let $\langle u, v\rangle$ denote the inner-product of the (d-dimensional) vectors u and v. Then, the Norm1 value of the vector v according to the system of coordinates (i.e., orthonormal basis) $b_1, ..., b_d$ equals $\sum_{k=1}^{d} |\langle v, b_k\rangle|$. The key technical observation is that, for an orthonormal basis $b_1, ..., b_d$ that is chosen uniformly at random, it holds that

$$\mathbf{E}_{b_1,...,b_d}\left[\sum_{k=1}^{d} |\langle v, b_k\rangle|\right] \;=\; d\cdot \mathbf{E}_{b_1}\left[|\langle v, b_1\rangle|\right] \;=\; d\cdot \|v\|\cdot \mathbf{E}_{b_1}\left[|\langle \overline{v}, b_1\rangle|\right], \tag{2}$$

where $\overline{v} = v/\|v\|$ is a unit vector in the direction of v. Furthermore, for any unit vector $u \in \mathsf{R}^d$, the value $\mathbf{E}_{b_1}\left[|\langle u, b_1\rangle|\right]$ is independent of the specific vector u, while b_1 is merely a uniformly distributed unit vector (in R^d). Thus, letting r denote a uniformly distributed unit vector, we define $\rho(d) \stackrel{\rm def}{=} \mathbf{E}_r\left[|\langle u, r\rangle|\right]$ and observe that

$$\mathbf{E}_{b_1}\left[|\langle v, b_1\rangle|\right] \;=\; \|v\|\cdot \rho(d)\,. \tag{3}$$

Moreover, a closed form expression for $\rho(d)$, which is linearly related to $1/\sqrt{d}$, is well-known (see Section 1.5).

Turning back to Eq. (3), we have $\|v\| = \mathbf{E}_r\left[|\langle v, r\rangle|\right]/\rho(d)$, where r is a random unit vector. It follows that

$$\sum_{i,j\in[n]} \|P_i - P_j\| \;=\; \frac{1}{\rho(d)}\cdot \mathbf{E}_r\left[\sum_{i,j\in[n]} |\langle P_i - P_j, r\rangle|\right]. \tag{4}$$

Noting that $|\langle P_i - P_j, r\rangle| = |\langle P_i, r\rangle - \langle P_j, r\rangle|$, this completes the randomized reduction of the d-dimensional case to the one-directional case; that is, the reduction selects a random unit vector r, and computes $\sum_{i,j\in[n]} |\langle P_i, r\rangle - \langle P_j, r\rangle|$. Note that we have obtained an *unbiased estimator*[1] for $\sum_{i,j\in[n]} \|P_i - P_j\|$. Furthermore, as shown in Section 1.5, this estimator is strongly concentrated around its expected value; in particular, the square root of the variance of this estimator is linearly related to its expectation. We thus obtain:

Theorem 1 *There exists a randomized algorithm that, given an approximation parameter $\epsilon > 0$ and points $P_1, ..., P_n \in \mathsf{R}^d$, runs for $\tilde{O}(\epsilon^{-2}\cdot|P_1, ..., P_n|)$-time and with probability at least $2/3$ outputs a value in the interval $[(1-\epsilon)\cdot A, (1+\epsilon)\cdot A]$, where $A = \sum_{i,j\in[n]} \|P_i - P_j\|/n^2$.*

Let us spell-out the algorithm asserted in Theorem 1 and complete its analysis. This algorithm consists of repeating the following procedure $O(\epsilon^{-2})$ times:

1. Uniformly select a unit vector $r \in \mathsf{R}^d$.

[1] A random variable X (e.g., the output of a randomized algorithm) is called an unbiased estimator of a value v if $\mathbf{E}[X] = v$.

2. For $i = 1, ..., n$, compute the projection $p_i = \langle P_i, r \rangle$.

3. Compute $\frac{1}{\rho(d)\cdot n^2} \sum_{i,j\in[n]} |p_i - p_j|$, by invoking the procedure described in Section 1.1 and using the value $\rho(d)$ computed as in Section 1.5.

The algorithm outputs the average of the values obtained in the various iterations. Step 2 can be implemented using $n \cdot d$ real (addition and multiplication) operations, whereas the complexity of Step 3 is dominated by sorting n real values.

The issues addressed next include the exact implementation of a single iteration (i.e., approximating real-value computations), which leads to an efficient derandomization, and providing an analysis of a single iteration (thus proving that $O(\epsilon^{-2})$ iterations suffice). Let us start with the latter.

1.4 Probabilistic analysis of a single iteration

Let us denote by X the random value computed by a single iteration, and let $Z = (\rho(d) \cdot n^2) \cdot X$. Recall that $Z = \sum_{i,j\in[n]} |\langle P_i, r\rangle - \langle P_j, r\rangle|$, which equals $\sum_{i,j\in[n]} |\langle P_i - P_j, r\rangle|$, where r is a uniformly distributed unit vector. Note that

$$\begin{aligned}
\mathbf{E}[Z] &= \mathbf{E}_r\left[\sum_{i,j\in[n]} |\langle P_i - P_j, r\rangle|\right] \\
&= \rho(d) \cdot \sum_{i,j\in[n]} \|P_i - P_j\|,
\end{aligned}$$

where the second equality relies on Eq. (4). This establishes the claim that *each iteration provides an unbiased estimator of* $\sum_{i,j\in[n]} \|P_i - P_j\|/n^2$. As usual, the usefulness of a single iteration is determined by the variance of the estimator. A simple upper-bound on the variance of Z may be obtained as follows

$$\begin{aligned}
\mathbf{V}[Z] &= \mathbf{V}_r\left[\sum_{i,j\in[n]} |\langle P_i - P_j, r\rangle|\right] \\
&\leq \mathbf{E}_r\left[\left(\sum_{i,j\in[n]} |\langle P_i - P_j, r\rangle|\right)^2\right] \\
&\leq \left(\sum_{i,j\in[n]} \|P_i - P_j\|\right)^2
\end{aligned}$$

where the second inequality uses the fact that $|\langle P_i - P_j, r\rangle| \leq \|P_i - P_j\|$ holds (for any unit vector r). This implies that $\mathbf{V}[Z] \leq \rho(d)^{-2} \cdot \mathbf{E}[Z]^2 = O(d \cdot \mathbf{E}[Z]^2)$. In Section 1.5, we will show that it actually holds that $\mathbf{V}[Z] = O(\mathbf{E}[Z]^2)$.

Applying Chebyshev's Inequality, it follows that the average value of t iterations (of the procedure) yields an $(1 \pm \epsilon)$-factor approximation with probability at least $1 - \frac{\mathbf{V}[Z]}{(\epsilon\cdot\mathbf{E}[Z])^2\cdot t}$. We stress that this holds even if the iterations are performed using random choices that are pairwise independent (rather than totally independent) among the various iterations. Thus, for $\mathbf{V}[Z] = O(\mathbf{E}[Z]^2)$, setting $t = O(\epsilon^{-2})$ will do.

1.5 On $\rho(d)$ and the related variance $\sigma^2(d)$

Recall that $\rho(d) \stackrel{\text{def}}{=} \mathbf{E}_r[|\langle u, r\rangle|]$, where u is an arbitrary unit vector (in R^d) and r is a uniformly distributed unit vector (in R^d). Analogously, we define the corresponding variance $\sigma^2(d) \stackrel{\text{def}}{=} \mathbf{V}_r[|\langle u, r\rangle|]$. Note that both $\rho(d)$ and $\sigma^2(d)$ are actually independent of the specific vector u.

Theorem 2 (folklore): *$\rho(d) = \frac{1}{(d-1)\cdot A_{d-2}}$, where $A_0 = \pi/2$, $A_1 = 1$, and $A_k = \frac{k-1}{k} \cdot A_{k-2}$.*

In particular, $\rho(2) = 2/\pi \approx 0.63661977$ and $\rho(3) = 1/2$. In general, $\rho(d) = \Theta(1/\sqrt{d})$. A proof of Theorem 2 can be found in [1, Sec. 3.6]. Using similar techniques (see [1, Sec. 3.7]), one may obtain

Theorem 3 (probably also folklore): *$\sigma^2(d) = O(1/d)$. Furthermore, for any two unit vectors $u_1, u_2 \in \mathsf{R}^d$ and for a uniformly distributed unit vector $r \in \mathsf{R}^d$, it holds that $\mathbf{E}_r[|\langle u_1, r\rangle| \cdot |\langle u_2, r\rangle|] = O(1/d)$.*

In fact, the furthermore clause follows from $\sigma^2(d) = O(1/d)$ (and $\rho(d)^2 = O(1/d)$) by using the Cauchy-Schwartz Inequality.[2]

Improved bound for the variance of Z. Recalling that $Z = \sum_{i,j\in[n]} |\langle P_i - P_j, r\rangle|$ and using Theorem 3, we have

$$
\begin{aligned}
\mathbf{V}[Z] \;\leq\; \mathbf{E}[Z^2] &= \mathbf{E}_r\left[\sum_{i_1,j_1,i_2,j_2\in[n]} |\langle P_{i_1} - P_{j_1}, r\rangle| \cdot |\langle P_{i_2} - P_{j_2}, r\rangle|\right] \\
&= O(1/d) \cdot \sum_{i_1,j_1,i_2,j_2\in[n]} \|P_{i_1} - P_{j_1}\| \cdot \|P_{i_2} - P_{j_2}\| \\
&= O(1/d) \cdot \left(\sum_{i,j\in[n]} \|P_i - P_j\|\right)^2 .
\end{aligned}
$$

Recalling that $\mathbf{E}[Z] = \rho(d) \cdot \sum_{i,j\in[n]} \|P_i - P_j\|$, it follows that $\mathbf{V}[Z] = O(1/d) \cdot (\mathbf{E}[Z]/\rho(d))^2$. Using $\rho(d) = \Omega(1/\sqrt{d})$, we conclude that $\mathbf{V}[Z] = O(\mathbf{E}[Z]^2)$.

1.6 Implementation details (i.e., the required precision)

By inspecting the various operations of our algorithm, one may verify that it suffices to conduct all calculations with $O(\log(1/\epsilon))$ bits of precision (see [1] for details). In particular, this holds with respect to the selection of $r \in \mathsf{R}^d$, which is the only randomization that occurs in a single iteration. It follows that each iteration can be implemented using $O(d \cdot \log(1/\epsilon))$ coin tosses (i.e., $O(\log(1/\epsilon))$ bits of precision per each coordinate of r).

Using pairwise independent random choices in the different iterations, the entire algorithm can be implemented within randomness-complexity $m = O(d \cdot \log(1/\epsilon))$. It follows that a full-derandomization can be obtained at the cost of slowing down the algorithm by a factor of $2^m = \text{poly}(1/\epsilon)^d$.

[2] That is, $\mathbf{E}_r[|\langle u_1, r\rangle| \cdot |\langle u_2, r\rangle|] \leq \sqrt{\mathbf{E}_r[|\langle u_1, r\rangle|^2] \cdot \mathbf{E}_r[|\langle u_2, r\rangle|^2]} = \mathbf{E}_r[|\langle u_1, r\rangle|^2] = \sigma^2(d) + \rho(d)^2$.

1.7 Reflection

In retrospect, the foregoing algorithm is an incarnation of the "embedding paradigm" (i.e., the fact that the Euclidean metric can be embedded with little distortion in the ℓ_1-metric). The point, however, is that we do not require a good (i.e., small distortion) embedding of *all* pairwise distances, but rather a good (i.e., small distortion) embedding of the *average* pairwise distances. Note that insisting on a good embedding of all $\binom{n}{2}$ pairwise distances would require a logarithmic number of dimensions (rather than a constant number of dimensions, as implicit in our scheme for constant $\epsilon > 0$). Either way, *computing the average ℓ_1-metric distances in a d'-dimensional space reduces to d' computations of average distances in a one-dimensional space.*

2 General Metric and the Sampling Algorithm

The straightforward algorithm for approximating the average pairwise distances consists of selecting a random sample of m pairs of points and outputting the average distance between these pairs. This algorithm works for any metric space. The question is how large should its sample be; that is, how should m relate to the number of points, denoted n. Indeed, m should be proportional to $\mathbf{V}[Z]/\mathbf{E}[Z]^2$, where Z represents the result of a single "distance measurement" (i.e., the distance between a uniformly selected pair of points). Specifically, to obtain an $(1 \pm \epsilon)$-approximation of the average of all pairwise distances, it suffices to take $m = O(\mathbf{V}[Z]/(\epsilon \cdot \mathbf{E}[Z])^2)$. Thus, we first upper-bound the ratio $\mathbf{V}[Z]/\mathbf{E}[Z]^2$, showing that it is at most linear in the number of points (see Section 2.1). We later consider the question of derandomization (see Section 2.2).

2.1 The approximation provided by a random sample

We consider an arbitrary metric $(\delta_{i,j})_{i,j\in[n]}$ over n points, where $\delta_{i,j}$ denote the distance between the ith and jth point. Actually, we shall only use the fact that the metric is symmetric (i.e., for every $i, j \in [n]$ it holds that $\delta_{i,j} = \delta_{j,i}$) and satisfies the triangle inequality (i.e., for every $i, j, k \in [n]$ it holds that $\delta_{i,k} \leq \delta_{i,j} + \delta_{j,k}$). Recall that Z is a random variable representing the distance between a uniformly selected pair of points; that is, $Z = \delta_{i,j}$, where $(i, j) \in [n] \times [n]$ is uniformly distributed.

Proposition 4 *For Z as above, it holds that $\mathbf{V}[Z] = O(n \cdot \mathbf{E}[Z]^2)$.*

Proof: By an averaging argument, it follows that there exists a point c (which may be viewed as a "center") such that

$$\frac{1}{n} \cdot \sum_{j\in[n]} \delta_{c,j} \leq \frac{1}{n^2} \cdot \sum_{i,j\in[n]} \delta_{i,j} \, . \tag{5}$$

Using such a (center) point c, we upper-bound $\mathbf{E}[Z^2]$ as follows:

$$\begin{aligned}
\mathbf{E}[Z^2] &= \frac{1}{n^2} \cdot \sum_{i,j\in[n]} \delta_{i,j}^2 \\
&\leq \frac{1}{n^2} \cdot \sum_{i,j\in[n]} (\delta_{i,c} + \delta_{c,j})^2 \\
&\leq \frac{1}{n^2} \cdot \sum_{i,j\in[n]} \left(2\delta_{i,c}^2 + 2\delta_{c,j}^2\right)
\end{aligned}$$

$$= \frac{4n}{n^2} \cdot \sum_{j\in[n]} \delta_{c,j}^2$$

where the first inequality is due to the triangle inequality and the last equality uses the symmetry property. Thus, we have

$$\begin{aligned}
\mathbf{E}[Z^2] &\leq 4n \cdot \sum_{j\in[n]} \left(\frac{\delta_{c,j}}{n}\right)^2 \\
&\leq 4n \cdot \left(\sum_{j\in[n]} \frac{\delta_{c,j}}{n}\right)^2 \\
&\leq 4n \cdot \left(\sum_{i,j\in[n]} \frac{\delta_{i,j}}{n^2}\right)^2
\end{aligned}$$

where the last inequality is due to Eq. (5). Thus, we obtain $\mathbf{E}[Z^2] \leq 4n \cdot \mathbf{E}[Z]^2$, and the proposition follows (because $\mathbf{V}[Z] \leq \mathbf{E}[Z^2]$). ■

Tightness of the bound. To see that Proposition 4 is tight, consider the metric $(\delta_{i,j})_{i,j\in[n]}$ such that $\delta_{i,j} = 1$ if either $i = v \neq j$ or $j = v \neq i$ and $\delta_{i,j} = 0$ otherwise. (Note that this metric can be embedded on the line with v at the origin (i.e. location 0) and all other points co-located at 1.) In this case $\mathbf{E}[Z] = 2(n-1)/n^2 < 2/n$ while $\mathbf{E}[Z^2] = \mathbf{E}[Z]$, which means that $\frac{\mathbf{V}[Z]}{\mathbf{E}[Z]^2} = \frac{1}{\mathbf{E}[Z]} - 1 > \frac{n}{2} - 1$.

2.2 On the limits of derandomization

A "direct" derandomization of the sampling-based algorithm requires trying all pairs of points, which foils our aim of obtaining a sub-quadratic time algorithm. Still, one may ask whether a better derandomization exists. We stress that such a derandomization should work well for all possible metric spaces. The question is how well can a *fixed* sample of pairs (over $[n]$) approximate the average distance between all the pairs (of points) in *any* metric space (over n points). The corresponding notion is formulated as follows.

Definition 5 (universal approximator): *A multi-set of pairs $S \subseteq [n] \times [n]$ is called a* universal (L, U)-approximator *if for every metric $(\delta_{i,j})_{i,j\in[n]}$ it holds that*

$$L(n) \cdot n^{-2} \cdot \sum_{i,j\in[n]} \delta_{i,j} \leq |S|^{-1} \cdot \sum_{(i,j)\in S} \delta_{i,j} \leq U(n) \cdot n^{-2} \cdot \sum_{i,j\in[n]} \delta_{i,j}. \quad (6)$$

In such a case, we also say that S is a universal U/L-approximator.

Needless to say, $[n] \times [n]$ itself is a universal 1-approximator, but we seek universal approximators of almost linear size. We shall show an explicit construction (of almost linear size) that provides a logarithmic-factor approximation, and prove that this is the best possible.

We note that universal approximators can be represented as n-vertex directed graphs (possibly with parallel and anti-parallel edges). In some cases, we shall present universal approximators as undirected graphs, while actually meaning the corresponding directed graph obtained by replacing each undirected edge with a pair of anti-parallel directed edges.

2.2.1 A construction

For an integer parameter k, we shall consider the generalized k-dimensional hypercube having n vertices, which are viewed as k-ary sequences over $[n^{1/k}]$ such that two vertices are connected by an edges if and only if (as k-long sequences) they differ in one position. That is, the vertices $\langle\sigma_1,...,\sigma_k\rangle \in [n^{1/k}]^k$ and $\langle\tau_1,...,\tau_k\rangle$ are connected if and only if $|\{i \in [k] : \sigma_i \neq \tau_i\}| = 1$. In addition, we add k self-loops to each vertex, where each such edge corresponds to some $i \in [k]$. Thus, the degree of each vertex in this n-vertex graph equals $k \cdot n^{1/k}$. We shall show that this graph constitutes a universal $O(k)$-approximator.

Theorem 6 *The generalized k-dimensional n-vertex hypercube is a universal $(1/k, 2)$-approximator.*

In particular, the binary hypercube (i.e., $k = \log_2 n$) on n vertices constitutes a universal $O(\log n)$-approximator.

Proof: For every two vertices $u, v \in [n]$, we consider a canonical path of length k between u and v. This path, denoted $P_{u,v}$, corresponds to the sequence of vertices $w^{(0)},...,w^{(k)}$ such that $w^{(i)} = \langle\sigma_1,...,\sigma_{k-i},\tau_{k-i+1},...,\tau_k\rangle$, where $u = \langle\sigma_1,...,\sigma_k\rangle$ and $v = \langle\tau_1,...,\tau_k\rangle$. (Here is where we use the self-loops.) Below, we shall view these paths as sequences of edges (i.e., $P_{u,v}$ is viewed as the k-long sequence $(w^{(0)}, w^{(1)}),...,(w^{(k-1)}, w^{(k)})$). An important property of these canonical paths is that each edge appears on the same number of paths.

Letting E denoted the directed pairs of vertices that are connected by an edge, and using the triangle inequality and the said property of canonical paths, we note that

$$\begin{aligned}
n^{-2} \cdot \sum_{u,v\in[n]} \delta_{u,v} &\leq n^{-2} \cdot \sum_{u,v\in[n]} \sum_{(w,w')\in P_{u,v}} \delta_{w,w'} \\
&= n^{-2} \cdot \sum_{(w,w')\in E} |\{(u,v) \in [n]\times[n] : (w,w') \in P_{u,v}\}| \cdot \delta_{w,w'} \\
&= n^{-2} \cdot \sum_{(w,w')\in E} \frac{k\cdot n^2}{|E|} \cdot \delta_{w,w'}
\end{aligned}$$

which equals $k \cdot |E|^{-1} \cdot \sum_{(w,w')\in E} \delta_{w,w'}$. On the other hand, letting $\Gamma(u) = \{v : (u,v) \in E\}$, and using the triangle inequality and the regularity of the graph, we note that

$$\begin{aligned}
|E|^{-1} \cdot \sum_{(u,v)\in E} \delta_{u,v} &\leq |E|^{-1} \cdot \sum_{(u,v)\in E} n^{-1} \cdot \sum_{w\in[n]} (\delta_{u,w} + \delta_{w,v}) \\
&= |E|^{-1} n^{-1} \cdot \sum_{u,w\in[n]} \sum_{v\in\Gamma(u)} (\delta_{u,w} + \delta_{w,v}) \\
&= |E|^{-1} n^{-1} \cdot \left(\sum_{u,w\in[n]} |\Gamma(u)| \cdot \delta_{u,w} + \sum_{v,w\in[n]} |\Gamma^{-1}(v)| \cdot \delta_{u,w} \right) \\
&= |E|^{-1} n^{-1} \cdot 2 \cdot \frac{|E|}{n} \cdot \sum_{u,w\in[n]} \delta_{u,w}
\end{aligned}$$

which equals $2 \cdot n^{-2} \cdot \sum_{u,w\in[n]} \delta_{u,w}$. ■

2.2.2 A lower-bound

We now show that the construction provided in Theorem 6 is optimal. Indeed, our focus is on the case $k < \log_2 n$ (and actually, even $k = o(\log n)$).

Theorem 7 *A universal k-approximator for n points must have $\frac{n^{(k+1)/k}}{4k}$ edges.*

Proof: Let $G = ([n], E)$ be (a directed graph representing) a universal k-approximator. We first note that no vertex can have (out)degree exceeding $2k \cdot (|E|/n)$ (even when not counting self-loops). The reason being that if vertex v has a larger degree, denoted d_v, then we reach contradiction by considering the metric $(\delta_{i,j})_{i,j\in[n]}$ such that $\delta_{i,j} = 1$ if either $i = v \neq j$ or $j = v \neq i$ and $\delta_{i,j} = 0$ otherwise. (Note that this metric can be embedded on the line with v at the origin (i.e. location 0) and all other points co-located at 1.) In this case $n^{-2} \cdot \sum_{i,j\in[n]} \delta_{i,j} < 2/n$, whereas $|E|^{-1} \cdot \sum_{(i,j)\in E} \delta_{i,j} = |E|^{-1} \cdot d_v$ (which is greater than $2k/n$).

We now consider the metric induced by the graph G itself; that is, $\delta_{i,j}$ equals the distance between vertices i and j in the graph G. Clearly, $|E|^{-1} \cdot \sum_{(i,j)\in E} \delta_{i,j} = 1$, but (as we shall see) the average distance between pairs of vertices is much larger. Specifically, letting $d \leq 2k \cdot (|E|/n)$ denote the maximum degree of a vertex in G, we have

$$\begin{aligned} n^{-2} \cdot \sum_{u,v\in[n]} \delta_{u,v} &\geq \min_{u\in[n]} \{n^{-1} \cdot \sum_{v\in[n]} \delta_{u,v}\} \\ &\geq n^{-1} \cdot \sum_{i=0}^{t} d^i \cdot i \end{aligned}$$

where t is the smallest integer such that $\sum_{i=0}^{t} d^i \geq n$, which implies $t > \log_d((1 - d^{-1}) \cdot n) \approx \frac{\ln n - d^{-1}}{\ln d}$. Thus, $n^{-2} \cdot \sum_{u,v\in[n]} \delta_{u,v}$ is lower-bounded by $\frac{d^t}{n} \cdot t \geq (1 - d^{-1}) \cdot t > (1 - 2d^{-1}) \cdot \frac{\ln n}{\ln d}$ which must be at most at most k (because otherwise G cannot be a universal k-approximator). Using $(1-2d^{-1}) \cdot \frac{\ln n}{\ln d} \leq k$ it follows that $d' \stackrel{\text{def}}{=} d^{1/(1-2d^{-1})} \geq n^{1/k}$, which (using $d' < 2d$) implies $2d > n^{1/k}$. Finally, using $d \leq 2k \cdot (|E|/n)$, we get $|E| \geq \frac{dn}{2k} > \frac{n^{(k+1)/k}}{4k}$. ■

2.3 On the limits of derandomization, revisited

Note that while the first part of the proof of Theorem 7 (i.e., bounding the maximum degree in terms of the average-degree) uses an Euclidean metric, the main part of the proof refers to a graph metric (which may not have a Euclidean embedding). Thus, Theorem 7 does not rule out the existence of sparse graphs that provide good approximations for points in a Euclidean space.

Definition 8 (universal approximator, restricted): *A multi-set of pairs $S \subseteq [n] \times [n]$ is called a* (L, U)-approximator for the class $\mathcal{M}$ *if Eq. (6) holds for any n-point metric of the class $\mathcal{M}$.*

Needless to say, any universal (L, U)-approximator is (L, U)-approximator for the Euclidean metric, but the converse does not necessarily hold. Indeed, we shall see that approximators for the Euclidean metric can have much fewer edges than univeral approximators (for any metric).

Theorem 9 *For every constant $\epsilon > 0$, there exists a efficiently constructible $(1 + \epsilon)$-approximator for the Euclidean metric that has $O(n/\epsilon^2)$ edges. Furthermore, such approximators can be constructed in linear-time.*

Theorem 9 follows by reducing the general case of (high-dimensional) Euclidean metric to the line-metric (i.e., one-dimensional Euclidean metric) and presenting an approximator for the latter metric.

Proposition 10 *Suppose that $S \subseteq [n] \times [n]$ is an f-approximator for the line-metric. Then S constitutes an f-approximator for the Euclidean metric.*

Proof: Considering a d-dimentional Euclidean space with points $P_1, ..., P_n \in \mathsf{R}^d$, we let r denote a uniformly distributed unit vector in R^d. By Eq. (3), for every vector $v \in \mathsf{R}^d$ it holds that $\mathbf{E}_r[|\langle v, r\rangle|] = \rho(d) \cdot \|v\|$. Thus, for every $i, j \in [n]$ it holds that $\|P_i - P_j\| = \rho(d)^{-1} \cdot \mathbf{E}_r[|\langle P_i, r\rangle - \langle P_i, r\rangle|]$ and so

$$\begin{aligned}
\sum_{(i,j)\in S} \|P_i - P_j\| &= \rho(d)^{-1} \cdot \mathbf{E}_r\left[\sum_{(i,j)\in S} |\langle P_i, r\rangle - \langle P_i, r\rangle|\right] \\
\sum_{i,j\in[n]} \|P_i - P_j\| &= \rho(d)^{-1} \cdot \mathbf{E}_r\left[\sum_{i,j\in[n]} |\langle P_i, r\rangle - \langle P_i, r\rangle|\right]
\end{aligned}$$

The proposition follows by applying the hypothesis to (each value of r in) the r.h.s of each of the foregoing equalities. ■

Strong expanders. Good approximators for the line-metric are provided by the following notion of graph expansion. We say that the (undirected) graph $G = ([n], E)$ is a $(1 - \epsilon)$-strong expander if for every $S \subset [n]$ it holds that

$$\frac{|E(S, [n] \setminus S)|}{|E|} = (1 \pm \epsilon) \cdot \frac{|S| \cdot (n - |S|)}{n^2/2} \tag{7}$$

where $E(V_1, V_2) \stackrel{\text{def}}{=} \{\{u, v\} \in E : u \in V_1 \wedge v \in V_2\}$. As we shall see (in Proposition 12), sufficiently good expanders (i.e., having relative eigenvalue bound $\epsilon/2$) are strong expanders (i.e., $(1-\epsilon)$-strong expander). We first establish the connection between strong expanders and good approximators for the line-metric.

Proposition 11 *Suppose that the graph $G = ([n], E)$ is a $(1 - \epsilon)$-strong expander. Then it yields a $(1 + \epsilon)$-approximator for the line-metric.*

The sufficient condition regarding G is also necessary (e.g., for any cut $(S, [n] \setminus S)$, consider the points $p_1, ..., p_n \in \mathsf{R}$ such that $p_i = 0$ if $i \in S$ and $p_i = 1$ otherwise).

Proof: For any sequence of points $p_1, ..., p_n \in \mathsf{R}$, consider the "sorting permutation" $\pi : [n] \to [n]$ such that for every $i \in [n - 1]$ it holds that $p_{\pi(i)} \le p_{\pi(i+1)}$. By counting the contribution of each "line segment" $[p_{\pi(i)}, p_{\pi(i+1)}]$ to $\sum_{i,j\in[n]} |p_i - p_j|$, we get

$$\sum_{i,j\in[n]} |p_i - p_j| = \sum_{i=1}^{n-1} 2i \cdot (n - i) \cdot (p_{\pi(i+1)} - p_{\pi(i)}) \tag{8}$$

Similarly, for $S_i = \{\pi(1), ..., \pi(i)\}$, we have

$$\sum_{i,j:\{i,j\}\in E} |p_i - p_j| = \sum_{i=1}^{n-1} 2|E(S_i, [n] \setminus S_i)| \cdot (p_{\pi(i+1)} - p_{\pi(i)}) \tag{9}$$

Using the proposition's hypothesis, we have for every $i \in [n-1]$,

$$\frac{2 \cdot |E(S_i, [n] \setminus S_i)|}{2 \cdot |E|} = (1 \pm \epsilon) \cdot \frac{2 \cdot i \cdot (n-i)}{n^2} \tag{10}$$

and the proposition follows by combining Eq. (8)–(10). ■

Proposition 12 *Suppose that the graph $G = ([n], E)$ is a d-regular graph with a second eigenvalue bound $\lambda < d$. Then G is a $(1 - (2\lambda/d))$-strong expander.*

Thus any family of constructible $O(\epsilon^{-2})$-regular Ramanujan graphs (e.g., [4]) yields a constructible family of $(1-\epsilon)$-strong expanders. Furthermore, such graphs can be constructed in almost linear time (i.e., each edge can be determined by a constant number of arithmetic operations in a field of size smaller than n).

Proof: The claim follows from the Expander Mixing Lemma, which refers to any two sets $A, B \subseteq [n]$, and asserts that

$$\left| \frac{|E(A,B)|}{d \cdot n} - \rho(A) \cdot \rho(B) \right| \leq \frac{\lambda}{d} \cdot \sqrt{\rho(A) \cdot \rho(B)} \tag{11}$$

where $\rho(S) = |S|/n$ for every set $S \subseteq [n]$. Applying the lemma to the special case of $A = B$, we infer that $|E(A,A)|$ resides in the interval $[\rho(A) \cdot d \cdot |A| \pm \lambda \cdot |A|]$. Assuming, without loss of generality, that $|A| \leq n/2$, we conclude that $|E(A, [n] \setminus A)|/(d \cdot n)$ resides in $[\rho(A) \cdot (1 - \rho(A)) \pm (\lambda/d) \cdot \rho(A)]$, which is a sub-interval of $[(1 \pm (2\lambda/d)) \cdot \rho(A) \cdot (1 - \rho(A))]$ (because $1 - \rho(A) \geq 1/2$). ■

Conclusion. The foregoing three propositions imply the *existence of* (efficiently constructible) *$O(\epsilon^{-2})$-regular graphs are $(1-\epsilon)$-approximators for the Euclidean metric.* It is easy to see that the argument extends to the ℓ_1-metric. We note that combining the foregoing fact with the proof of Theorem 7 it follows that most constant-degree graphs cannot be embedded in a Euclidean space (resp., ℓ_1-metric) without incurring logarithmic distortion. Details follow.

Recall that a metric (e.g., a graph metric) on n points, denoted $(\delta_{i,j})_{i,j \in [n]}$ is said to be embedded in a Euclidean space with distortion ρ if the distance between points i and j in the embedding is at least $\delta_{i,j}$ and at most $\rho \cdot \delta_{i,j}$. Thus, if a graph satisfying Proposition 12 can be embedded in a Euclidean space with distortion ρ, then this graph constitutes a $(1+\epsilon) \cdot \rho$-approximator of the graph metric induced by itself. But then the proof of Theorem 7 implies that for $k = (1+\epsilon) \cdot \rho$ it holds that $O(\epsilon^{-2}) \geq n^{1/k}/4k$. Actually, since the graph in question is regular, we may skip the first part of the said proof and use $d = 2|E|/n$ (rather than $d = 2k|E|/n$), thus obtaining $O(\epsilon^{-2}) \geq n^{1/k}/2$. Using any fixed $\epsilon > 0$, it follows that $k = \Omega(\log n)$.

Acknowledgments

We are grateful to Bernard Chazelle, Dan Halperin, Robi Krauthgamer, Nati Linial, David Peleg, Gideon Schechtman, and Avi Wigderson for helpful discussions.

References

[1] K. Barhum. Approximating Averages of Geometrical and Combinatorial Quantities. M.Sc. Thesis, Weizmann Institute of Science, February 2007. Available from `http://www.wisdom.weizmann.ac.il/~oded/msc-kb.html`

[2] U. Feige. On sums of independent random variables with unbounded variance, and estimating the average degree in a graph. In *Proc. of the 36th STOC*, pages 594–603, 2004.

[3] O. Goldreich and D. Ron. Approximating Average Parameters of Graphs. In *Proc. of the 10th RANDOM*, Springer LNCS 4110, pages 363–374, 2006.

[4] A. Lubotzky, R. Phillips, and P. Sarnak. Ramanujan Graphs. *Combinatorica*, Vol. 8, pages 261–277, 1988.