

# Chapter 2

## Low-Distortion Embeddings

Informally speaking, a *metric space* is a set of points for which we have a reasonable notion of distance between any two points. In this chapter, we will study questions of the following kind: How well can one kind of metric space (which we think of as “complicated”) be embedded into another one (which we think of as “simple”) in such a way that the distances are approximately preserved? Our presentation is mostly based on Chapter 15 of Matoušek’s book [7] and on [6].

### 2.1 Metric Spaces and Normed Spaces

This section contains the formal definitions of metric spaces and normed spaces. You may prefer to jump ahead to the following sections and refer back to the first one and look up the definitions when they are needed later.

**Definition 2.1** (Metrics and Metric Spaces). *Suppose we have a set  $X$  (finite or infinite) and a function  $\rho : X \times X \rightarrow \mathbb{R}$  such that the following conditions hold for all  $x, y, z \in X$ :*

1.  $\rho(x, y) \geq 0$ , and  $\rho(x, y) = 0$  iff  $x = y$ ,
2.  $\rho(x, y) = \rho(y, x)$ ,
3.  $\rho(x, z) \leq \rho(x, y) + \rho(y, z)$ .

*Then  $\rho$  is called a metric and the pair  $(X, \rho)$  is called a metric space. When the metric  $\rho$  is understood from the context, we will sometimes just speak of the metric space  $X$  and suppress  $\rho$  from the notation.*

As we will see below, checking the 3rd property, called the *triangle inequality*, is usually the nontrivial part when determining whether a given function  $\rho$  is a metric or not.

### Some Examples.

1. The most familiar case is  $X = \mathbb{R}^d$  with the Euclidean distance  $\rho(x, y) = \|x - y\|_2$ . In fact, this is an example of a *normed space* (see below).
2. The *Hamming cube*, i.e., the set  $X = \{0, 1\}^d$  with the hamming distance  $\rho_H(x, y) = |\{i : x_i \neq y_i\}|$ , which we also already encountered in Chapter 1.
3. If  $X$  is the set of all finite strings over some finite alphabet  $\Sigma$ , we can define the *edit distance*  $\rho_{\text{edit}}(x, y)$  between two strings  $x, y \in X$  to be the minimal number of substitutions (replace one letter by another one), deletions, or insertions needed to transform  $x$  into  $y$  (or vice versa).
4. If  $G = (V, E)$  is a graph, we can define the *shortest path metric* on the vertices: For any two vertices  $v, w \in V$ , we define their distance  $\rho(v, w)$  as the length of a shortest path in the graph connecting  $v$  and  $w$ .

Note that if we view the discrete cube as a graph in the usual way, i.e.,  $x, y \in \{0, 1\}^d$  are connected by an edge if they differ in exactly one coordinate, then the resulting shortest path metric is exactly the Hamming metric.

Usually, one assumes that the graph underlying the shortest path metric is finite, but in fact one can allow infinite vertex sets as long as any two vertices are connected by some finite path. Then also the edit distance is a special case of a shortest path metric, namely on the following graph: The vertices are all finite strings over the given alphabet, and two strings  $x$  and  $y$  are connected by an edge iff  $x$  can be obtained from  $y$  by exactly one deletion, one insertion, or one substitution.

If  $X$  is a vector space, then it is quite natural to consider metrics that are in some sense compatible with the operations allowed in a vector space, addition and scalar multiplication. A metric on  $X$  is called *translation-invariant* if the distance from  $x$  to  $y$  is the same as the distance from  $x + z$  to  $y + z$ , for all  $x, y, z \in X$ . For such a translation-invariant metric, it is in fact enough to know the distance from every point to some fixed point, say the origin  $\mathbf{0}$ ; this already determines the metric, because  $\rho(x, y) = \rho(x - y, \mathbf{0})$ .

If we additionally require that the metric be scaling-sensitive, then we arrive at the definition of a norm:

**Definition 2.2** (Norms and Normed Spaces). *Let  $X$  be a real vector space.<sup>1</sup> A norm on  $X$  is a function  $\|\cdot\| : X \rightarrow \mathbb{R}$  with the following properties:*

1.  $\|x\| \geq 0$  for all  $x \in X$ , and  $\|x\| = 0$  iff  $x = \mathbf{0}$ .

---

<sup>1</sup> $X$  could be infinite-dimensional, although we will mostly be concerned with  $X \cong \mathbb{R}^d$ ; the definition also makes sense for complex vector spaces.

2.  $\|\lambda x\| = |\lambda|\|x\|$  for all  $x \in X$  and  $\lambda \in \mathbb{R}$ .

3.  $\|x + y\| \leq \|x\| + \|y\|$  for all  $x, y \in X$ .

**Remark 2.3.** Any norm defines translation-invariant metric on  $X$  by setting  $\rho(x, y) := \|x - y\|$  (check this!). Conversely, if we start with a metric  $\rho$  with these additional properties, we can define a norm by  $\|x\| := \rho(x, 0)$ .

A viewpoint that is very useful for understanding a norm is by looking at its “unit ball”  $B_1(\|\cdot\|) := \{x \in X : \|x\| \leq 1\}$ . Observe that the unit ball determines a norm in the sense that

$$\|x\| = \inf\{\lambda \in \mathbb{R}_+ : \frac{1}{\lambda}x \in B_1(\|\cdot\|)\}.$$

The unit ball of a norm is a centrally symmetric convex set that is bounded in the sense that it does not contain any infinite ray. If the dimension of  $X$  is finite, then the unit ball is also compact, but in infinite-dimensional spaces, this need not be the case.

**Example 2.4 ( $p$ -Norms).** For a real number  $1 \leq p < \infty$ , the  $p$ -norm  $\|\cdot\|_p$  on  $\mathbb{R}^d$  is defined by

$$\|x\|_p := \sqrt[p]{\sum_{i=1}^d |x_i|^p}, \quad x = (x_1, \dots, x_d) \in \mathbb{R}^d.$$

The usual Euclidean norm is the special case  $p = 2$ . Observe that as  $p$  grows larger and larger, the influence of coordinates of large absolute value increases. The definition of the  $p$ -norm is extended to the limit case  $p = \infty$  by setting

$$\|x\|_\infty := \max_{1 \leq i \leq d} |x_i|.$$

For  $1 \leq p \leq \infty$ , we write  $\ell_p^d$  for the space  $\mathbb{R}^d$  equipped with the  $p$ -norm  $\|\cdot\|_p$ .

As remarked before, when proving that these functions  $\|\cdot\|_p$  are indeed norms, checking the triangle inequality is the interesting part. Except for the cases  $p = 1$  and  $p = \infty$ , this is nontrivial and boils down to the following:

**Fact 2.5 (Hölder’s Inequality).** Let  $1 \leq p, q \leq \infty$  such that  $1/p + 1/q = 1$  (where we interpret  $1/\infty = 0$ ); the numbers  $p$  and  $q$  are called conjugate exponents. Observe, for instance, that 1 and  $\infty$  are conjugate exponents, and that 2 is its own conjugate exponent. Then, for all  $x, y \in \mathbb{R}^d$ ,

$$|\langle x, y \rangle| \leq \|x\|_p \|y\|_q.$$

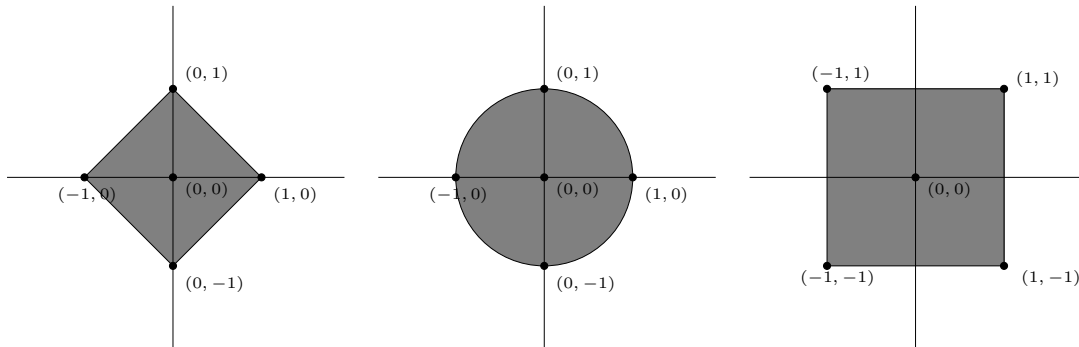


Figure 2.1: The unit balls in  $\ell_1^2$ ,  $\ell_2^2$ , and  $\ell_\infty^2$

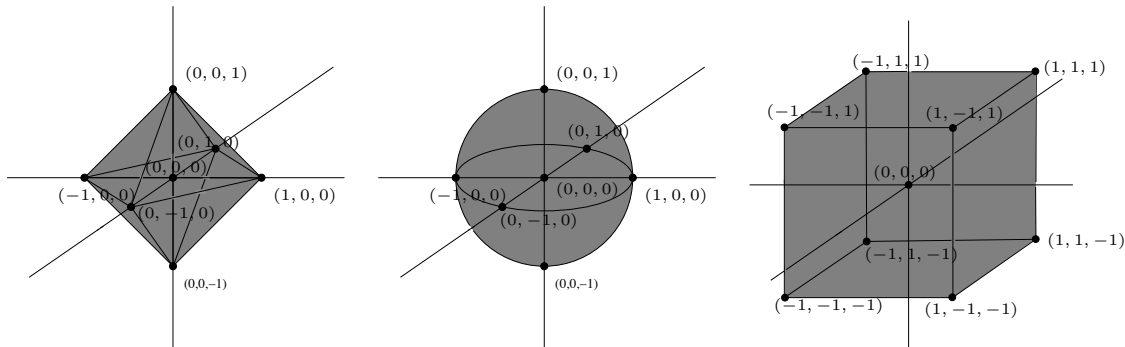


Figure 2.2: The unit balls in  $\ell_1^3$ ,  $\ell_2^3$ , and  $\ell_\infty^3$ .

The most important  $p$ -norms are the 1-norm, the 2-norm, and the  $\infty$ -norm. Figures 2.1 and 2.2 show their respective unit balls in dimensions 2 and 3.

Sometimes, it is also convenient to extend these definitions to infinite sequences: For a sequence  $\mathbf{x} = (x_i)_{i=1}^\infty$  of real (or complex) numbers, we define  $\|\mathbf{x}\|_p := \sqrt[p]{\sum_{i=1}^\infty |x_i|^p}$  if  $1 \leq p < \infty$ , and  $\|\mathbf{x}\|_\infty := \sup_i |x_i|$ , and

$$\ell_p := \{\mathbf{x} = (x_i)_{i=1}^\infty : \|\mathbf{x}\|_p < \infty\}.$$

## 2.2 Low-Distortion Embeddings

A metric  $\rho$  on an  $n$ -element set  $X$  can be specified by writing down all the  $\binom{n}{2}$  pairwise distances (in a symmetric  $n \times n$ -matrix, say). By contrast, if  $X$  is an  $n$ -elements subset of some space  $\ell_p^d$ , it suffices to write down the  $d \cdot n$  coordinates of the points, and we can still compute the pairwise distances in  $O(d)$  steps. If  $d$  is much smaller than  $n$ , this is more space efficient. Moreover, it is generally quite difficult to discern any patterns or structure in a large table of numbers, while it is often much easier to do so in a low-dimensional normed space.

As an often quoted example, think of a large number  $n$  of different strains of bacteria; the pairwise distances could be given by comparing the DNA (the DNA of each strain can be viewed as a word over the 4-letter alphabet  $\{A, C, G, T\}$ , and we could take the edit-distance) or some other biologically interesting measure of (dis)similarity. We might be interested in detecting large clusters of similar bacteria. Or maybe we would like to store the information about the numerous bacteria that we already know in such a way that when somebody claims that they found a new one, we can quickly decide whether it is similar to one in our database.

Such tasks can be generally quite difficult. In contrast, in low-dimensional Euclidean (and other normed) spaces, efficient geometric algorithms and data structure (for clustering or nearest-neighbor queries) are available. This raises the following general question: Given a “complicated” metric space  $X$ , can we represent it in some “simpler” space  $Y$ ? For instance, in an ideal situation, we might be able to *isometrically embed*  $(X, \rho)$  into the Euclidean plane  $\ell_2^2$ , i.e., to assign to every point  $x \in X$  a point  $f(x) \in \mathbb{R}^2$  such that for any two points  $x, y \in X$ , their distance  $\rho(x, y)$  is the same as the Euclidean distance  $\|f(x) - f(y)\|_2$  between their image points. In this case, we could immediately visualize  $X$  and see clusters and other structures right away.

It is easy to see that isometric embeddings are generally too much to hope for. For instance, consider the two 4-point metric spaces defined by the shortest-path metrics on the two graphs in Figure 2.3. Neither of these two metric

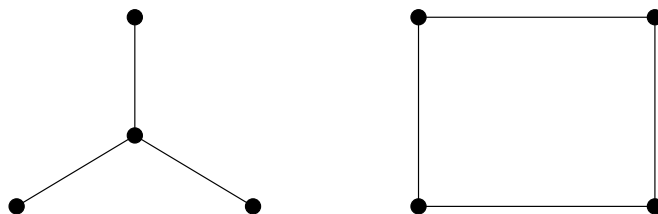


Figure 2.3: Not isometrically embeddable into Euclidean space.

spaces is isometrically embeddable into the Euclidean plane (or indeed into any  $\ell_2^d$ ; this is Exercise 10).

However, for many purposes it suffices if we can find a mapping such that the distances are *approximately preserved*, and this leads to a rich theory, of which we will just scratch the surface.

**Definition 2.6.** Let  $(X, \rho)$  and  $(Y, \sigma)$  be metric spaces, and let  $D \geq 1$  be a real number. A map  $f : X \rightarrow Y$  is said to have distortion at most  $D$  if the following holds: There exists a real number  $r > 0$  (a “scaling factor”) such that for all  $x, y \in X$ ,

$$r \cdot \rho(x, y) \leq \sigma(f(x), f(y)) \leq D \cdot r \cdot \rho(x, y).$$

The distortion of  $f$  is defined as the infimum of all  $D$  such that the above holds; if no such  $D$  exists, then the distortion of  $f$  is  $\infty$ .

We will focus on target spaces  $Y$  that are normed spaces. In this case, we can always take the scaling factor  $r$  to be equal to 1 (by replacing  $f$  by  $\frac{1}{r} \cdot f$ , if necessary), even though it will still sometimes be convenient to work with a different  $r$ .

As an example, consider the metric space defined by the second graph (the square) in Figure 2.3. If we embed it into the Euclidean plane by mapping the nodes of the graph to the vertices of the unit square,  $(0, 0)$ ,  $(1, 0)$ ,  $(0, 1)$ , and  $(1, 1)$  in the obvious way, then the resulting map has distortion  $\sqrt{2}$ : The distances between adjacent vertices are preserved exactly, and for each pair of antipodal vertices, their distance in the graph equals two, while the Euclidean distance equals  $\sqrt{2}$ . More generally, if we take the natural embedding of the Hamming cube into Euclidean space, i.e., if we consider the set  $\{0, 1\}^d$  as a subset of  $\mathbb{R}^d$ , then this inclusion map has distortion  $\sqrt{d}$ : this follows from the observation that for  $x, y \in \{0, 1\}^d$ , we have  $\rho_H(x, y) = \|x - y\|$ . Moreover, by the Cauchy-Schwarz inequality, for any  $z \in \mathbb{R}^d$ ,

$$\|z\|_1 = \sum_{i=1}^d 1 \cdot |z_i| = \langle \mathbf{1}, (|z_i|)_{i=1}^d \rangle \leq \|\mathbf{1}\|_2 \|z\|_2 = \sqrt{d} \|z\|_2.$$

Applying this to the difference vectors  $z = x - y$ , we see that  $\|x - y\|_2 \leq \rho_H(x, y) = \|x - y\|_1 \leq \sqrt{d} \|x - y\|_2$  for all  $x, y \in \{0, 1\}^d$ , and the upper bound is attained, for instance, for  $x = \mathbf{0}$  and  $y = \mathbf{1}$ .

A concept that is closely related to distortion is that of *Lipschitz maps*.

**Definition 2.7.** Let  $(X, \rho)$  and  $(Y, \sigma)$  be metric spaces, and let  $C > 0$  be a real number. A map  $f : X \rightarrow Y$  is said to be  $C$ -Lipschitz if  $\sigma(f(x), f(y)) \leq C \cdot \rho(x, y)$  for all  $x, y \in X$ . The Lipschitz constant  $\|f\|_{\text{Lip}}$  of  $f$  is defined as the infimum over all  $C$  for which  $f$  is  $C$ -Lipschitz.<sup>2</sup>

Note that a map  $f$  with distortion at most  $D < \infty$  is necessarily injective, and by replacing  $Y$  by  $f(X) \subseteq Y$ , we may assume that  $f$  is bijective. For a bijective map  $f : X \rightarrow Y$  between metric spaces, the distortion of  $f$  equals  $\|f\|_{\text{Lip}} \cdot \|f^{-1}\|_{\text{Lip}}$  (Exercise 11). For this reason, maps of finite distortion are also often called *bi-Lipschitz*.

The study of low-distortion maps between metric spaces is a very active area of mathematics with numerous applications in computer science.

---

<sup>2</sup>The notation is no accident; if  $Y$  is a normed space, then the set of all functions  $f : X \rightarrow Y$  with finite Lipschitz constant forms a vector space, and  $\|f\|_{\text{Lip}}$  defines a norm on this space.

## 2.3 The Johnson-Lindenstrauss Flattening Lemma

Our first topic is that of *dimension reduction* for subsets of Euclidean space, i.e., on low-distortion embeddings from finite subsets of high-dimensional Euclidean spaces into low-dimensional ones. There are finite subsets of the  $d$ -dimensional Euclidean space  $\ell_2^d$  that cannot be isometrically embedded into any lower-dimensional Euclidean space; the most basic example is the set of vertices of a regular  $d$ -dimensional simplex (this is a set of  $d + 1$  points in  $\ell_2^d$  such that all the pairwise distances equal 1). The goal of this section is to prove the following theorem, which states that the dimension can be significantly reduced if we allow for a small distortion  $(1 + \varepsilon)$ .

**Theorem 2.8 (Johnson-Lindenstrauss Flattening Theorem).** *Let  $X$  be a set of  $n$  points in the  $d$ -dimensional Euclidean space  $\ell_2^d$ , and let  $0 < \varepsilon \leq 2$  be a parameter. Then there exists a map  $T : X \rightarrow \ell_2^k$  of distortion at most  $(1 + \varepsilon)$ , where<sup>3</sup>  $k = O(\varepsilon^{-2} \log n)$ . In fact,  $T$  can be taken as the restriction of a linear map  $\mathbb{R}^d \rightarrow \mathbb{R}^k$  to  $X$ .*

Thus, if we do not need to know the precise pairwise distances but can allow an error of  $\varepsilon = 5\%$ , then we can embed any  $n$ -point set in Euclidean space (for instance, the  $n$  vertices of the regular simplex, whose natural habitat is in dimension  $d = n - 1$ ) into dimension  $O(\log n)$ . As an immediate application, this implies a dramatic reduction in storage: To store  $n$  points in  $\ell_2^n$ , we need  $n^2$  numbers; similarly, we need to store  $\binom{n}{2}$  numbers for the exact pairwise distances. However, after dimension reduction, it suffices to store  $O(n \log n)$  numbers, and we can still reconstruct the pairwise distances up to an error of 5%.

It may be surprising at the first moment that the original dimension  $d$  does not appear in the bound for the target dimension  $k$ . The reason is that for the purposes of the theorem, we may assume that  $d < n$  by restricting our attention to the affine hull of the point set  $X$  (if  $|X| = n$ , then the affine hull  $\text{aff}(X)$  is of dimension at most  $n - 1$ ).

The currently best constant for the embedding dimension is  $(4 + o(1))$ , where  $o(1) \rightarrow 0$  as  $\varepsilon \rightarrow 0$ . We also remark that the bound for the embedding dimension is almost sharp, for a wide range of  $n$  and  $\varepsilon$ . The lower-bound example is the vertex set of a regular  $(n - 1)$ -dimensional simplex. This was shown by Alon and is the subject of Exercises 18, 19, and 20.

The idea for the proof of the Flattening Lemma is to take the map  $T$  at random according to a suitable probability distribution on the space of linear maps  $\mathbb{R}^d \rightarrow \mathbb{R}^k$ , or in other words, to choose a random matrix  $T \in \mathbb{R}^{k \times d}$ . The main technical step is then to show that for any *fixed* unit vector  $u \in \mathbb{R}^d$ ,

---

<sup>3</sup>A more precise formulation that includes all the quantifiers in the right order is the following: There exist constants  $C, \varepsilon_0 > 0$  such that for every  $0 < \varepsilon \leq \varepsilon_0$ , there is an  $n_0$  such that for all  $n \geq n_0$ , for all  $d \geq 1$ , for all  $k \geq C \cdot \varepsilon^{-2} \log n$  and for all  $n$ -element subsets  $X$  of  $\ell_2^d$ , there is a map  $g : X \rightarrow \mathbb{R}^k$  of distortion at most  $1 + \varepsilon$ .

$\|u\| = 1$ , the length of the image,  $\|Tu\|_2$  is sharply concentrated around its mean. The “classical” proof uses projections onto random  $k$ -dimensional subspaces. The proof we will discuss below shows that we can choose the entries  $T_{ij}$  of the matrix  $T$ , to be independently and identically distributed for any one from a large class of well-behaved probability distributions. For instance, we can take the  $T_{ij}$  to be independent standard normal random variables, or even  $\pm 1$ -random variables with  $\Pr[T_{ij} = +1] = \Pr[T_{ij} = -1] = 1/2$ , which is computationally very simple.

## 2.4 Measure Concentration for Sums of Independent Random Variables

In this section, we prove some technical results concerning random variables that are strongly concentrated around their expectation, and derive the Johnson-Lindenstrauss Theorem. All random variables will be real-valued.

The most basic estimate for the probability that a random variable differs from its expectation is

**Fact 2.9 (Markov’s Inequality).** *Let  $X \geq 0$  be a nonnegative random variable. Then*

$$\Pr[X \geq \lambda] \leq \frac{\mathbf{E}[X]}{\lambda}$$

for all  $\lambda > 0$ .

If  $X$  is a random variable (not necessarily nonnegative) for which  $\mathbf{E}[X]$  exists and is finite, then we also have estimate the probability that  $X$  deviates from its expectation in terms of the *variance*

$$\text{Var}[X] := \mathbf{E}[(X - \mathbf{E}[X])^2] = \mathbf{E}[X^2] - (\mathbf{E}[X])^2.$$

**Fact 2.10 (Chebyshev’s Inequality).** *If  $X$  is a random variable for which  $\mathbf{E}[X]$  exists and is finite, then*

$$\Pr[|X - \mathbf{E}[X]| \geq \lambda] \leq \frac{\text{Var}[X]}{\lambda^2}$$

for all  $\lambda > 0$ .

In Markov’s inequality, the estimate for the probability decreases linearly as  $\lambda \rightarrow \infty$ , and in Chebyshev’s Inequality, it decreases quadratically.

In this section, we consider random variables for which there is an exponential decrease. More precisely:

**Definition 2.11 (Subgaussian tails).** Let  $X$  be a random variable with  $\mathbf{E}[X] = 0$ . We say that  $X$  has a subgaussian upper tail if there exists a constant  $a > 0$  such that

$$\Pr[X > \lambda] \leq e^{-a\lambda^2} \quad (2.1)$$

for all  $\lambda > 0$ .

We say that  $X$  has a subgaussian upper tail up to  $\lambda_0$  if (2.1) holds for all  $0 < \lambda \leq \lambda_0$ . We say that  $X$  has a *subgaussian tail* if both  $X$  and  $-X$  have subgaussian upper tails.

**Examples 2.12.** 1. Let  $X$  be a standard normal random variable, i.e.,  $X$  has probability distribution function

$$\Pr[X \leq \lambda] = \Phi(\lambda) := \underbrace{\int_{-\infty}^{\lambda} \frac{1}{\sqrt{2\pi}} e^{-t^2/2} dt}_{=: \varphi(\lambda)}.$$

We have  $\mathbf{E}[X] = 0$  (in fact,  $X$  is symmetric about the origin) and  $\text{Var}[X] = 1$ . Note that  $-\varphi'(t) = t\varphi(t)$ . Thus,

$$\begin{aligned} \Pr[X > \lambda] &= 1 - \Phi(\lambda) = \int_{\lambda}^{\infty} \varphi(t) dt \\ &\leq \int_{\lambda}^{\infty} \underbrace{\frac{t}{\lambda}}_{\geq 1} \varphi(t) dt = -\frac{1}{\lambda} \int_{\lambda}^{\infty} \varphi'(t) dt = \frac{\varphi(\lambda)}{\lambda}. \end{aligned}$$

Thus,  $\Pr[X > \lambda] \leq \frac{1}{\sqrt{2\pi}\lambda} e^{-\lambda^2/2}$ . This is almost, but not quite, of the desired form. To remedy this, note first of all that by symmetry of the normal distribution,  $\Pr[X > \lambda] < 1/2$  for all  $\lambda > 0$ . For  $\lambda$  close to 0, more precisely, for  $0 < \lambda < \frac{1}{\sqrt{2\pi}}$ , we thus have  $\Pr[X > \lambda] < 1/2 < e^{-a_1\lambda^2}$  if we choose  $a_1 := \frac{\ln 2}{2\pi}$ , and for  $\lambda \geq 1/\sqrt{2\pi}$ , we have  $\Pr[X > \lambda] \leq e^{-a_2\lambda^2}$  with  $a_2 = 1/2$ , so the smaller of these two  $a$ 's works for all  $\lambda$ .

2. Let  $X$  be a balanced  $\pm 1$  random variable, i.e.,  $\Pr[X = +1] = \Pr[X = -1] = 1/2$ . Then

$$\begin{aligned} \Pr[X > \lambda] &= \frac{1}{2} \quad \text{for } 0 < \lambda < 1, \text{ and} \\ \Pr[X > \lambda] &= 0 \quad \text{for } \lambda \geq 1/2, \end{aligned}$$

and this can easily be brought into the desired form.

We will also need the following standard estimates for the exponential function:

$$1 + x \leq e^x \quad \text{for all } x \in \mathbb{R} \quad (2.2)$$

$$e^x \leq 1 + 2x \quad \text{for all } x \in [0, 1] \quad (2.3)$$

$$e^x \leq 1 + x + x^2 \quad \text{for all } x \leq 1 \quad (2.4)$$

$$\frac{e^x + e^{-x}}{2} \leq e^{x^2/2} \quad \text{for all } x \in \mathbb{R} \quad (2.5)$$

$$e^{-x} \leq 1/x \quad \text{for all } x > 0. \quad (2.6)$$

All of these are easy to prove using the power series expansion  $e^x = \sum_{k=0}^{\infty} x^k/k!$  for the exponential function.

We will show that if we choose the entries of the projection matrix  $T$  as independent random variables with a uniform subgaussian tail (this means that the same constant  $a$  works for all random variables), zero expectation and variance 1, then we get the following version of the Johnson-Lindenstrauss Flattening Theorem:

**Theorem 2.13 (Johnson-Lindenstrauss Theorem, technical version).** *Let  $n \in \mathbb{N}$ ,  $0 < \varepsilon \leq 1/2$ , and  $0 < \delta < 1$ . If we set  $k := C \cdot \varepsilon^{-2} \ln(2/\delta)$ , where  $C$  is a suitable constant to be determined later, and if  $X_{ij}$ ,  $1 \leq i \leq k$ ,  $1 \leq j \leq n$  are independent random variables with  $\mathbf{E}[X_{ij}] = 0$ ,  $\text{Var}[X_{ij}] = \mathbf{E}[X_{ij}^2] = 1$  and a uniform subgaussian tail<sup>4</sup> then the matrix  $T := \frac{1}{\sqrt{k}}[X_{ij}]$  has the following property:*

*For any fixed unit vector  $u \in \mathbb{R}^n$ ,  $\|u\|_2 = 1 = \sum_j u_j^2$  the length of the image vector  $Tu \in \mathbb{R}^k$  is close to 1 with high probability in the sense that*

$$\Pr[1 - \varepsilon \leq \|Tu\|_2 \leq 1 + \varepsilon] \geq 1 - \delta.$$

Let us first see how this implies Theorem 2.8: Suppose we are given an  $n$ -point set  $S \subseteq \mathbb{R}^d$  and an error parameter  $0 < \varepsilon \leq 2$ . The goal is to exhibit a map  $T : X \rightarrow \mathbb{R}^k$  of distortion at most  $(1 + \varepsilon)$ , where  $k = O(\varepsilon^{-2} \ln n)$ . As remarked above, w.l.o.g. we may assume that  $d = n$ . Let  $\tilde{\varepsilon} := \varepsilon/4 \leq 1/2$  and  $\delta := 1/n^2$ . Let  $k = O(\tilde{\varepsilon}^{-2} \ln(2/\delta)) = O(\varepsilon^{-2} \ln n)$  as in Theorem 2.13 and let  $\tilde{T} : \mathbb{R}^n \rightarrow \mathbb{R}^k$  be the corresponding random linear map. Fix a pair  $x, y \in S$ ,  $x \neq y$ , of distinct points in  $S$  and let  $u := \frac{1}{\|x-y\|_2}(x-y)$  be their difference, rescaled to length 1. By Theorem 2.13, with probability at least  $1 - 1/n^2$ , we have

$$1 - \tilde{\varepsilon} \leq \|\tilde{T}u\|_2 \leq 1 + \tilde{\varepsilon}.$$

Moreover, by linearity,  $\tilde{T}u = \frac{1}{\|x-y\|_2}(\tilde{T}x - \tilde{T}y)$ . Thus, multiplying the inequality by  $\|x-y\|_2$  and dividing by  $1 - \tilde{\varepsilon}$ , we see that with probability at least  $1 - 1/n^2$ ,

---

<sup>4</sup>i.e.,  $\Pr[X_{ij} > \lambda] < e^{-a\lambda^2}$  for all  $\lambda > 0$ , with the same constant  $a$  for all  $X_{ij}$ ; the constant  $C$  in the definition of  $k$  depends on  $a$

the rescaled linear map  $T := \frac{1}{\varepsilon} \tilde{T}$  is “good” for  $x$  and  $y$  in the sense that

$$\|x - y\|_2 \leq \|Tx - Ty\|_2 \leq \frac{1 + \tilde{\varepsilon}}{1 - \tilde{\varepsilon}} \|x - y\|_2 \leq (1 + \varepsilon) \|x - y\|_2,$$

where we are using the estimate  $\frac{1 + \tilde{\varepsilon}}{1 - \tilde{\varepsilon}} \leq 1 + 4\tilde{\varepsilon}$  for  $0 < \tilde{\varepsilon} \leq 1/2$ . Since there are  $\binom{n}{2}$  pairs in total, with probability at least  $1 - \binom{n}{2}/n^2 > 1/2$ , the map  $T$  is good for all pairs simultaneously. This shows that Theorem 2.13 implies Theorem 2.8.

In order to prove the former, we need a few auxiliary lemmata concerning the concentration of sums of independent random variables. The first three lemmata generalize the proof of the well-known Chernov bounds.

**Lemma 2.14 (Moment generating function and subgaussian tails).** *Let  $X$  be a random variable with  $\mathbf{E}[X] = 0$ . If  $\mathbf{E}[e^{\lambda X}] \leq e^{C\lambda^2}$  for some constant  $C$  and all  $\lambda > 0$  then  $X$  has a subgaussian upper tail. If  $\mathbf{E}[e^{\lambda X}] \leq e^{C\lambda^2}$  for all  $\lambda \in (0, \lambda_0]$  then  $X$  has a subgaussian upper tail up to  $2C\lambda_0$ .*

We leave the proof as an exercise. Here is a partial converse (under the additional condition  $\text{Var}[X] = 1$ ):

**Lemma 2.15.** *Let  $X$  be a random variable with  $\mathbf{E}[X] = 0$  and  $\text{Var}[X] = \mathbf{E}[X^2] = 1$ . If  $X$  has a subgaussian upper tail then  $\mathbf{E}[e^{\lambda X}] \leq e^{C\lambda^2}$  for all  $\lambda > 0$ , where  $C$  depends on the constant  $a$  in the tail estimate.*

*Proof.* We formulate the proof in terms of *conditional expectations*.<sup>5</sup> Recall that for (finitely or countably many) mutually exclusive events  $A_j$  with  $\sum_j \Pr[A_j] = 1$  and a random variable  $Y$ , we have  $\mathbf{E}[Y] = \sum_j \mathbf{E}[Y|A_j] \Pr[A_j]$ . Let  $\lambda > 0$  and apply this to the random variable  $Y = e^{\lambda X}$ . We split the expectation according to the events  $A_0 = \{\lambda X \leq 1\}$  and  $A_j = \{j < \lambda X \leq j + 1\}$ ,  $j = 1, 2, 3, \dots$ :

$$\begin{aligned} \mathbf{E}[e^{\lambda X}] &= \mathbf{E}[e^{\lambda X} | X \leq 1/\lambda] \Pr[X \leq 1/\lambda] \\ &\quad + \sum_{j=1}^{\infty} \underbrace{\mathbf{E}[e^{\lambda X} | j < \lambda X \leq j + 1]}_{\leq e^{j+1} \leq e^{2j}} \underbrace{\Pr[j < \lambda X \leq j + 1]}_{\leq \Pr[X > j/\lambda] \leq e^{-aj^2/\lambda^2}}, \end{aligned}$$

where  $a$  is the constant in the definition of the subgaussian tail. For the first conditional expectation, we use (2.4):  $e^{\lambda X} \leq 1 + \lambda X + \lambda^2 X^2$  for  $\lambda X \leq 1$ , hence

$$\underbrace{\mathbf{E}[e^{\lambda X} | X \leq 1/\lambda]}_{\leq \mathbf{E}[1 + \lambda X + \lambda^2 X^2 | \lambda X \leq 1]} \Pr[\lambda X \leq 1] \leq \mathbf{E}[1 + \lambda X + \lambda^2 X^2] = 1 + \lambda \underbrace{\mathbf{E}[X]}_{=0} + \lambda^2 \underbrace{\mathbf{E}[X^2]}_{=1} = 1 + \lambda^2.$$

<sup>5</sup>Equivalently, we can phrase it in terms of the distribution function  $F(t) := \Pr[X \leq t]$  and Lebesgue *Lebesgue-Stieltjes integration*  $\int g(t) dF(t) = \mathbf{E}[g(X)]$ , as we did in class (where we assume that the function  $g$  is measurable, so that the integral makes sense). Lebesgue-Stieltjes integration and conditional expectations are really just two formal ways of saying the same thing.

For the second part, we have to estimate the series

$$\sum_{j=1}^{\infty} e^{j(2-aj/\lambda^2)}. \quad (2.7)$$

If  $\lambda \leq \sqrt{a}/2$ , we have

$$2 - aj/\lambda^2 \leq -a/(2\lambda^2) \leq -2$$

for all  $j$ . Therefore, in this case our series is dominated from above by the convergent geometric series

$$\sum_{j=1}^{\infty} \left( e^{-a/(2\lambda^2)} \right)^j = \frac{e^{-a/(2\lambda^2)}}{1 - e^{-a/(2\lambda^2)}} < 2e^{-a/(2\lambda^2)} \leq 4\lambda^2/a,$$

where we used the crude upper bound  $e^{-a/(2\lambda^2)} \leq e^{-2} < 1/2$  in the second-to-last step and (2.6) in the last one. Therefore,  $\mathbf{E}[e^{\lambda X}] \leq 1 + \lambda^2 + 4\lambda^2/a \leq e^{\lambda^2(1+4/a)}$ , as desired.

On the other hand, if  $\lambda > \sqrt{a}/2$  then the largest terms of the series are those with  $j$  close to  $\lambda^2/a$ , and the series in (2.7) is at most  $e^{O(\lambda^2)}$ ; more precisely, for  $j > 3\lambda^2/a$ , we have  $2 - ja/\lambda^2 < -1$ , and the tail of the series satisfies

$$\sum_{j>3\lambda^2/a} e^{j(2-ja/\lambda^2)} \leq \sum_{j=0}^{\infty} e^{-j} = \frac{1}{1 - e^{-1}} < 2 < 8\lambda^2/a.$$

Moreover, the function  $y \mapsto y(2 - ya/\lambda^2)$  is maximized (for  $y > 0$ ) at  $y = \lambda^2/a$ . Thus,  $e^{j(2-ja/\lambda^2)} \leq e^{\frac{\lambda^2}{a}(2-\frac{\lambda^2}{a}a/\lambda^2)} = e^{\lambda^2/a}$  for all  $j$ . Therefore, the initial part of the series satisfies

$$\sum_{j=1}^{\lfloor 3\lambda^2/a \rfloor} e^{j(2-ja/\lambda^2)} \leq (3\lambda^2/a)e^{\lambda^2/a} \leq e^{4\lambda^2/a}.$$

Hence, also in this case  $\mathbf{E}[e^{\lambda X}] \leq 1 + \lambda^2 + 8\lambda^2/a + e^{4\lambda^2/a} = e^{O(\lambda^2)}$ .  $\square$

**Lemma 2.16 (Generalized Chernov Bound).** *Let  $X_1, \dots, X_n$  be independent random variables with a uniform subgaussian tail and satisfying  $\mathbf{E}[X_i] = 0$  and  $\text{Var}[X_i] = \mathbf{E}[X_i^2] = 1$  for all  $i$ . If  $u_1, \dots, u_n$  are real coefficients with  $u_1^2 + \dots + u_n^2 = 1$  (in other words, the vector  $u = (u_1, \dots, u_n) \in \mathbb{R}^n$  satisfies  $\|u\|_2 = 1$ ) then the sum*

$$Y := u_1 X_1 + \dots + u_n X_n$$

*satisfies  $\mathbf{E}[Y] = 0$ ,  $\text{Var}[Y] = 1$ , and has a subgaussian tail.*

*Proof.* We will use the fact that for a product  $Z = \prod_i Z_i$  of independent random variables,  $\mathbf{E}[Z] = \prod_i \mathbf{E}[Z_i]$ . The first statement,  $\mathbf{E}[Y] = 0$ , is immediate by linearity of expectation. Consequently,

$$\text{Var}[Y] = \mathbf{E}[Y^2] = \underbrace{\sum_i u_i^2 \mathbf{E}[X_i^2]}_{=1} + \underbrace{\sum_{1 \leq i \neq j \leq n} u_i u_j \mathbf{E}[X_i X_j]}_{=0} = 1$$

since  $\mathbf{E}[X_i X_j] = \mathbf{E}[X_i] \mathbf{E}[X_j] = 0$  for  $i \neq j$ , by independence. Next, by Lemma 2.15, we have  $\mathbf{E}[e^{\lambda X_i}] \leq e^{C\lambda^2}$  for all  $\lambda > 0$  and all  $i$ , with a uniform constant  $C$ . Thus,

$$\mathbf{E}[e^{\lambda Y}] = \prod_{i=1}^n \mathbf{E}[e^{\lambda u_i X_i}] \leq e^{C\lambda^2(u_1^2 + \dots + u_n^2)},$$

so  $Y$  has a subgaussian upper tail by Lemma 2.14. The subgaussian lower tail follows symmetrically by applying the same argument to  $-Y$ .  $\square$

In the special case that the  $X_i$  are independent balanced  $\pm 1$  random variables and  $u_i = 1/\sqrt{n}$ , we retrieve the standard

**Corollary 2.17 (Chernov Bound).** *Let  $X_1, \dots, X_n$  be independent random variables with  $\Pr[X_i = +1] = \Pr[X_i = -1] = 1/2$ , and let  $X := \frac{1}{\sqrt{n}} \sum_{i=1}^n X_i$ . Then  $X$  has a subgaussian tail, i.e.,*

$$\Pr\left[\sum_i X_i > \lambda\sqrt{n}\right] = \Pr\left[\sum_i X_i < -\lambda\sqrt{n}\right] \leq e^{-a\lambda^2}$$

for some constant  $a > 0$  independent of  $n$  (as a matter of fact,  $a = 1/4$  works).

The following lemma captures the essence of Theorem 2.13:

**Lemma 2.18.** *Let  $Y_1, \dots, Y_k$  be independent random variables with  $\mathbf{E}[Y_i] = 0$  and  $\text{Var}[Y_i] = 1$  for all  $i$  and with a uniform subgaussian tail, i.e.,  $\Pr[Y_i > \lambda] \leq e^{-a\lambda^2}$  for all  $\lambda > 0$ , with the same constant  $a$  for all  $Y_i$ . Then*

$$Z := \frac{1}{\sqrt{k}} (Y_1^2 + \dots + Y_k^2 - k)$$

has a subgaussian tail up to  $\sqrt{k}$ .

Let us first see how this lemma implies the theorem:

*Proof that Lemma 2.18 implies Theorem 2.13.* Let  $u = (u_1, \dots, u_n) \in \mathbb{R}^n$  be a fixed unit vector, i.e.,  $\sum_i u_i^2 = 1$ , and let  $k = C\varepsilon^{-2} \ln(2/\delta)$ , the random variables  $X_{ij}$ ,

and the matrix  $T = \frac{1}{\sqrt{k}}[X_{ij}] \in \mathbb{R}^{k \times n}$  be as in the theorem. For each  $1 \leq i \leq k$ , the  $i^{\text{th}}$  coordinate of  $Tu$  is equal to

$$(Tu)_i = \frac{1}{\sqrt{k}} \underbrace{\sum_{j=1}^n X_{ij} u_j}_{=: Y_i}.$$

By the generalized Chernov bound of Lemma 2.16, the random variable  $Y_i$  satisfies  $\mathbf{E}[Y_i] = 0$ ,  $\text{Var}[Y_i] = 1$ , and has a subgaussian tail. Moreover,

$$\|Tu\|_2^2 - 1 = \frac{1}{k}(Y_1^2 + \dots + Y_k^2 - k) = \frac{1}{\sqrt{k}}Z,$$

where  $Z$  is the random variable considered in Lemma 2.18 above. By that lemma, we have

$$\begin{aligned} \Pr[\|Tu\|_2 \geq 1 + \varepsilon] &\leq \Pr[\|Tu\|_2^2 \geq 1 + 2\varepsilon] = \Pr[Z \geq 2\varepsilon\sqrt{k}] \\ &\leq e^{-a(2\varepsilon\sqrt{k})^2} \\ &= e^{-4a\varepsilon^2 C \varepsilon^{-2} \ln(2/\delta)} = e^{-4aC \ln(2/\delta)}, \end{aligned}$$

where we used the fact that  $\varepsilon \leq 1/2$  to be in the admissible range for the subgaussian tail. Thus, if we choose  $C = \frac{1}{2a}$ , we get  $\Pr[\|Tu\|_2 \geq 1 + \varepsilon] \leq \delta/2$ . The computation for  $\Pr[\|Tu\|_2 \leq 1 - \varepsilon] \leq \delta/2$  is analogous.  $\square$

Thus, it remains to prove Lemma 2.18. To this end, we need another auxiliary lemma similar to that used in the proof of the generalized Chernov bound.

**Lemma 2.19.** *let  $Y$  be a random variable with  $\mathbf{E}[Y] = 0$ ,  $\text{Var}[Y] = 1$  and with a subgaussian tail. Then there exist constants  $C \geq 1/2$  and  $\lambda_0 > 0$  such that*

$$\mathbf{E} \left[ e^{\lambda(Y^2-1)} \right] \leq e^{C\lambda^2} \quad \text{and} \quad \mathbf{E} \left[ e^{\lambda(1-Y^2)} \right] \leq e^{C\lambda^2}$$

for  $0 \leq \lambda \leq \lambda_0$ .

*Proof.* We start with the first inequality. Observe that  $\mathbf{E}[Y^4]$  is bounded from above by a constant; this follows from the subgaussian tail of  $Y$  and from Lemma 2.15, since  $t^4 = O(e^t + e^{-t})$  for all  $t \in \mathbb{R}$ .

First note that  $\mathbf{E}[e^{\lambda(Y^2-1)}] = e^{-\lambda} \mathbf{E}[e^{\lambda Y^2}]$ . As in the proof of Lemma 2.15, we split the expectation into conditional expectations according to whether  $\lambda Y^2 \leq 1$  or  $\lambda Y^2 > 1$ :

$$\begin{aligned} \mathbf{E}[e^{\lambda Y^2}] &= \underbrace{\mathbf{E}[e^{\lambda Y^2} | \lambda Y^2 \leq 1]}_{\leq 1 + \lambda + O(\lambda^2)} \cdot \Pr[\lambda Y^2 \leq 1] \\ &\leq \mathbf{E}[1 + \lambda Y^2 + \lambda^2 Y^4] \\ &\leq 1 + \lambda + O(\lambda^2) \\ &\quad + \sum_{k=1}^{\infty} \underbrace{\mathbf{E}[e^{\lambda Y^2} | k < \lambda Y^2 \leq k+1]}_{\leq e^{k+1} \leq e^{2k}} \cdot \underbrace{\Pr[k+1 \leq \lambda Y^2 > k]}_{\leq e^{-ak/\lambda}} \end{aligned}$$

We define  $\lambda_0$  (whose existence is claimed in the lemma) as  $\lambda_0 := a/4$ . Then, for  $0 < \lambda \leq \lambda_0$ , we have  $k(2 - a/\lambda) \leq -ka/(2\lambda)$ . Thus, the series is dominated by a convergent geometric series  $\sum_{k=1}^{\infty} c^k = \frac{c}{1-c}$ , where  $c = e^{-\frac{a}{2\lambda}}$ . As in the proof of Lemma 2.14, this is bounded from above by  $e^{-\Omega(1/\lambda)} = O(\lambda^2)$ . Thus, we obtain  $\mathbf{E}[e^{\lambda Y^2}] = 1 + \lambda + O(\lambda^2) \leq e^{\lambda + O(\lambda^2)}$ , and hence  $\mathbf{E}[e^{\lambda(Y^2-1)}] = e^{O(\lambda^2)}$ , as desired. For the second expected value, we can directly use (2.4) and obtain  $\mathbf{E}[e^{-\lambda Y^2}] \leq \mathbf{E}[1 - \lambda Y^2 + \lambda^2 Y^4] = 1 - \lambda + O(\lambda^2) \leq e^{-\lambda + O(\lambda^2)}$ , and hence  $\mathbf{E}[e^{\lambda(1-Y^2)}] = e^{O(\lambda^2)}$ .  $\square$

*Proof of Lemma 2.18.* Let  $Z = \frac{1}{\sqrt{k}}$  and  $\lambda_0$  in the preceding lemma. Then, for  $0 < \lambda \leq \sqrt{k}\lambda_0$ , we have

$$\mathbf{E}[e^{\lambda Z}] = \mathbf{E}\left[e^{(\lambda/\sqrt{k})(Y_1^2 + \dots + Y_k^2 - k)}\right] = \prod_{i=1}^k \mathbf{E}\left[e^{(\lambda/\sqrt{k})(Y_i^2 - 1)}\right] \leq \left(e^{C\lambda^2/k}\right)^k = e^{C\lambda^2}.$$

Thus, by Lemma 2.14,  $Z$  has a subgaussian upper tail up to  $2C\sqrt{k} \geq \sqrt{k}$ . The argument for the lower tail is analogous.  $\square$

## 2.5 Semidefinite Programming

Semidefinite programming means optimizing a linear function (of the matrix entries) over the set of all positive definite  $n \times n$ -matrices subject to finitely many linear inequalities. It is powerful technique in the design of approximation algorithms. Here, we will illustrate it by means of one application, the Goemans-Williamson approximation algorithm for the MAXCUT problem. Our discussion is based on Chapter 7 of the lecture notes [2] and on the survey article [5]. We first recall the definition of positive definite matrices:

**Definition 2.20.** Let  $A = [a_{ij}]_{i,j=1}^n \in \mathbb{R}^{n \times n}$  be a symmetric matrix (i.e.,  $A^T = A$ , or in other words,  $a_{ij} = a_{ji}$  for all  $i, j$ ).  $A$  is called positive semidefinite, denoted by  $A \succcurlyeq 0$ , if the quadratic form defined by  $A$  is nonnegative, i.e., if  $x^T A x \geq 0$  for all  $x \in \mathbb{R}^n$  (where  $\cdot^T$  denotes the transpose, here turning a column vector into a row vector, so that matrix multiplication makes sense).<sup>6</sup>

We recall from linear algebra that a symmetric real  $n \times n$ -matrix  $A$  has  $n$  real eigenvalues  $\lambda_1, \dots, \lambda_n \in \mathbb{R}$  with corresponding eigenvectors  $u_1, \dots, u_n$ ,  $Au_i = \lambda_i u_i$ , that form an orthonormal basis of  $\mathbb{R}^n$ , i.e., the  $u_i$  are pairwise orthogonal, of euclidean length  $\|u_i\|_2 = 1$ . Equivalently, there is an orthogonal matrix  $U \in \mathbb{R}^{n \times n}$ , i.e., one that satisfies  $U^T U = I$ , such that  $U^T A U$  is the diagonal matrix with entries  $\lambda_1, \dots, \lambda_n$ .

<sup>6</sup>Observe that the requirement that  $A$  be symmetric is not a real restriction: For any square matrix  $A$ ,  $x^T A x$  is a real number, so  $x^T A x = (x^T A x)^T = x^T A^T x$ . Thus, if  $x^T A x = x^T \tilde{A} x$ , where  $\tilde{A} := \frac{1}{2}(A + A^T)$  is the symmetrization of  $A$

**Lemma 2.21.** Let  $A = [a_{ij}] \in \mathbb{R}^{n \times n}$  be a symmetric matrix. Then the following statements are equivalent:

- (i)  $A \succcurlyeq 0$ .
- (ii) All eigenvalues of  $A$  are nonnegative:  $\lambda_1, \dots, \lambda_n \geq 0$ .
- (iii) There is a matrix  $B \in \mathbb{R}^{d \times n}$ , for some  $d \geq 0$ , such that  $A = B^T B$ . In other words, there are vectors  $b_1, \dots, b_n \in \mathbb{R}^d$  (the columns of  $B$ ) such that  $a_{ij} = \langle b_i, b_j \rangle$  for all  $i, j$ .

We leave the proof of this lemma as an exercise. The matrix  $B$  in (iii) is called a *Gram matrix* for  $A$ , and the factorization  $A = B^T B$  a *Gram decomposition*. Observe that  $n$  vectors always span a linear subspace of dimension at most  $n$ , so we may assume  $d \leq n$ . A special kind of Gram decomposition is the *Cholesky decomposition* that is studied in numerical analysis: For every symmetric positive semidefinite matrix  $A \in \mathbb{R}^{n \times n}$ , there is a *unique* upper triangular matrix  $U \in \mathbb{R}^{n \times n}$  (i.e.,  $u_{ij} = 0$  for  $j > i$ ) such that  $A = U^T U$ . If the input matrix  $A$  is positive definite and has rational entries, the matrix  $U$  can be computed efficiently, i.e., in time polynomial in  $n$  and in the bit sizes of the entries of  $A$ . To be more exact, the entries of the matrix  $U$  may be irrational<sup>7</sup> but can be computed efficiently to arbitrary precision (see [8], for instance).

A *semidefinite program* (SDP) is an optimization problem of the following form: Optimize (maximize or minimize) a linear function  $\sum_{i,j} c_{ij} x_{ij}$  of the entries of the matrix  $X = [x_{ij}]$  over all symmetric positive semidefinite matrices  $X \in \mathbb{R}^{n \times n}$  that satisfy a finite number of linear inequalities in the matrix entries, i.e., finitely many inequalities  $\sum_{ij} b_{ijk} x_{ij} \leq \beta_k$ ,  $1 \leq k \leq m$ . That is, the numbers  $c_{ij}$ ,  $b_{ijk}$  and  $\beta_k$  are given and specify the semidefinite program. There are many possible equivalent forms to write an SDP, for instance as

$$\begin{aligned} & \text{minimize} && c^T x = c_1 x_1 + \dots + c_m x_m \\ & \text{subject to} && x_1 A_1 + \dots + x_m A_m - B \succcurlyeq 0, \end{aligned} \tag{2.8}$$

where  $A_1, \dots, A_m, B$  are given symmetric  $n \times n$  matrices and  $c \in \mathbb{R}^m$  is a given vector. (It is a useful exercise to convince yourself that every SDP can be brought into this form.) Observe that in the special case that  $A_1, \dots, A_m, B$  are diagonal matrices, the SDP is equivalent to a linear program.

**Fact 2.22.** *Semidefinite programs can be solved efficiently. More precisely, there is an algorithm that, given an SDP with rational coefficients  $b_{ijk}$  and  $\beta_k$  and an error*

---

<sup>7</sup>For instance, if  $A$  is the diagonal matrix with 2's on the diagonal, then  $U$  is a diagonal matrix with  $\sqrt{2}$  on the diagonal. The standard Cholesky decomposition algorithm requires the computation of  $n$  square roots and  $O(n^3)$  additions and multiplications.

parameter  $\varepsilon > 0$ , computes a rational positive definite  $n \times n$ -matrix  $Y$  such that  $\sum_{ij} b_{ijk} y_{ij} \leq \beta_k$  for all  $k$  and

$$\sum_{i,j} c_{ij} y_{ij} \leq \text{opt} + \varepsilon,$$

where  $\text{opt}$  is the optimal value of the SDP, i.e., the infimum of  $\sum_{i,j} c_{ij} x_{ij}$  over all positive definite matrixes  $X$  satisfying all constraints of the SDP.<sup>8</sup> The runtime of the algorithm is polynomial in  $n$ ,  $\log(1/\varepsilon)$ , and in the bitsizes of the coefficients  $\beta_{ijk}$  and  $b_k$ .

We will not prove this or discuss the algorithm (the relevant buzzword is “interior point methods”); instead, we will discuss an application that exemplifies how semidefinite programming can be used to design approximation algorithms for hard optimization problems.

## An Approximation Algorithm for MAXCUT

A *cut* in a graph  $G = (V, E)$  is a partition  $V = S \cup (V \setminus S)$  of the vertex set into two disjoint parts. The *size* or *weight* of the cut is the total number of edges across the cut, i.e.,  $|E(S, V \setminus S)|$ , where  $E(S, V \setminus S) = \{\{u, v\} \in E : u \in S, v \in V \setminus S\}$ .

The MAXCUT problem is the combinatorial optimization problem of finding a cut of maximum size in a given graph  $G$ . The decision version of the problem<sup>9</sup> is NP-complete. Here, we present a (randomized) polynomial-time approximation algorithm, due to Goemans and Williamson, that uses semidefinite programming and achieves an (expected) *approximation ratio* of roughly 0.878. That is, given a graph  $G$ , the algorithm computes a cut, making some random choices along the way, such that the expected number of edges across the computed cut is at least 0.878 times the maximum size of any cut in  $G$ .

It will be convenient to work with the following “arithmetization” of the problem. Suppose we are given a graph  $G = (V, E)$ . Without loss of generality,

---

<sup>8</sup>To be precise, for this statement to be correct, one has to assume that the SDP is feasible (i.e., that there exists at least one positive semidefinite matrix  $X$  satisfying all constraints) and bounded, i.e., that  $\text{opt} > -\infty$ . Both assumptions will be satisfied in our applications, and in general, an arbitrary SDP can be transformed into one that is feasible and bounded, and such that solving the new SDP allows one to decide whether the original one was unbounded or infeasible. We also remark that, even for a feasible and bounded SDP, the optimum need not be attained. For example, the if we want to minimize  $x_1$  subject to the condition that

$$\begin{bmatrix} x_1 & 1 \\ 1 & x_2 \end{bmatrix} \succcurlyeq 0$$

(this is an SDP in the form (2.8)), the semidefiniteness boils down to  $x_1, x_2 \geq 0$  and  $x_1 x_2 \geq 1$ . Thus,  $\text{opt} = 0$ , but this is not attained.

<sup>9</sup>Given a graph  $G$  and an integer  $k$ , decide whether there is a cut in  $G$  of size at least  $k$ .

assume that  $V = \{1, 2, \dots, n\}$ . Define coefficients  $a_{ij}$ ,  $1 \leq i < j \leq n$  by  $a_{ij} = 1$  if  $\{i, j\} \in E$  and  $a_{ij} = 0$  if  $\{i, j\} \notin E$ . Then we can rephrase the problem of finding a maximum cut in  $G$  as the following integer quadratic optimization problem:

$$\text{maximize } \sum_{1 \leq i < j \leq n} a_{ij} \frac{1 - x_i x_j}{2}, \quad x_i \in \{+1, -1\}, 1 \leq i \leq n \quad (2.9)$$

To see that the optimal solution to (2.9) corresponds to a maximum cut in  $G$ , observe that there is a bijection between subsets  $S \subseteq V$  and  $\pm 1$ -vectors  $x = (x_1, \dots, x_n) \in \{+1, -1\}^n$ , namely  $x \leftrightarrow S = \{i : x_i = +1\}$ . Moreover,  $\frac{1 - x_i x_j}{2}$  is equal to 1 if  $x_i$  and  $x_j$  have different signs, and 0 otherwise. Thus, under the above bijection,

$$\sum_{1 \leq i < j \leq n} a_{ij} \frac{1 - x_i x_j}{2} = \sum_{\{i, j\} \in E} \frac{1 - x_i x_j}{2} = |E(S, V \setminus S)|.$$

Observe that this immediately gives a very simple randomized approximation algorithm: Choose random signs  $X_i$  independently at random with  $\Pr[X_i = +1] = \Pr[X_i = -1] = 1/2$ . Then  $\Pr[X_i X_j = -1] = \Pr[X_i X_j = 1] = 1/2$ , by independence, and therefore the expected size of the resulting random cut equals

$$\mathbf{E} \left[ \sum_{1 \leq i < j \leq n} a_{ij} \frac{1 - X_i X_j}{2} \right] = \sum_{1 \leq i < j \leq n} a_{ij} \frac{1 - \mathbf{E}[X_i X_j]}{2} = \frac{|E|}{2}.$$

Since  $|E(S, V \setminus S)| \leq |E|$  for any  $S$ , this expectation is at least  $1/2$  times the size of a maximum cut, i.e., we have a randomized approximation algorithm with expected approximation ratio at least  $1/2$ .

We will now use geometry and semidefinite programming to get a better approximation ratio. One way of motivating the improved algorithm is as follows: We can think of the choices  $+1$  and  $-1$  as 1-dimensional unit length vectors. We relax (2.9) by allowing unit vectors that live in higher dimensions (and replacing the product of real numbers by the scalar product):

$$\text{maximize } \sum_{1 \leq i < j \leq n} a_{ij} \frac{1 - \langle u_i, u_j \rangle}{2}, \quad u_i \in \mathbb{S}^{n-1}, 1 \leq i \leq n, \quad (2.10)$$

where  $\mathbb{S}^{n-1} = \{u \in \mathbb{R}^n : \|u\|_2 = 1\}$  is the set of unit vectors in  $\mathbb{R}^n$ , i.e., the  $(n - 1)$ -dimensional unit sphere centered at the origin. Observe that there is no need to consider unit vectors in higher dimensions,<sup>10</sup> since the linear

---

<sup>10</sup>One could consider, however, an intermediate problem, where the  $u_i$  are required to lie in some  $\mathbb{S}^{d-1}$ ,  $1 < d < n$ . For  $d = 1$ , the problem is equivalent to MAXCUT, and hence NP-hard. Lovász [5] mentions that he is not aware of any specific hardness results, but that he expects the problem to be NP-hard for any fixed  $d$ .

span of any set of  $n$  vectors is always of dimension at most  $n$ . Also note that the optimum of (2.10) is always at least as large as that of (2.9), since we can interpret any collection of  $n$  signs  $x_1, \dots, x_n \in \{+1, -1\}$  as unit vectors  $u_i = (x_i, 0, \dots, 0) \in \mathbb{S}^{n-1}$ ,  $1 \leq i \leq n$  by simply filling up the remaining coordinates with zeros.

We remark that one can think of (2.10) in physical terms as the problem of placing the vertices of the graph on the unit sphere  $\mathbb{S}^{n-1}$ , where there is a repulsive force between adjacent vertices that grows linearly with the distance (for instance, the edges could be springs that want to expand), and we want to minimize the “energy”  $\mathcal{E} = -\sum_{\{i,j\} \in E} \|u_i - u_j\|^2$ .

For our purposes, however, it is more relevant that (2.10) can be rewritten as semidefinite program. Namely, given  $u_1, \dots, u_n \in \mathbb{S}^{n-1}$ , we can consider the positive semidefinite matrix  $Y = [y_{ij}]$  given by  $y_{ij} = \langle u_i, u_j \rangle$  (in particular with diagonal entries  $y_{ii} = 1$ ), and conversely, given such a matrix  $Y$ , we can retrieve unit vectors  $u_i$  by computing a Gram decomposition for  $Y$ . The Gram decomposition is not unique, but we only care about the scalar products among the  $u_i$ , and these are uniquely determined by  $Y$ .

Thus, we get the following SDP:

$$\text{maximize } \sum_{1 \leq i < j \leq n} a_{ij} \frac{1 - y_{ij}}{2}, \quad Y \succeq 0, y_{ii} = 1, 1 \leq i \leq n. \quad (2.11)$$

We could ignore the constant terms in the objective function (since they sum up to  $|E|/2$ ) and instead consider the problem of minimizing the linear function  $\sum_{i < j} a_{ij} y_{ij}$  to bring the SDP closer to a standard form, but this is immaterial. Note also that the constraints are exactly of the required form: Positive semidefiniteness of the matrix  $Y$ , together with some linear inequalities on the matrix entries ( $y_{ii} = 1$  can be expressed by two inequalities  $y_{ii} \leq 1$  and  $y_{ii} \geq 1$ ).

As mentioned above, we can solve this SDP and compute a Gram decomposition  $Y = U^T U$  in polynomial time (more precisely, we can do this to arbitrary precision, an issue which we are going to ignore in the sequel). The constraints  $y_{ii} = 1$  imply that the columns  $u_i$  of the matrix  $U$  are unit vectors<sup>11</sup>, and so an optimal solution of (2.11) yields an optimal solution of (2.10).

Thus, we have found a set of unit vectors  $u_i \in \mathbb{S}^{n-1}$  that is “spread out” in the sense that on average, adjacent  $u_i$  and  $u_j$  are “far” from each other. How do we use this to obtain a large cut in the original graph? The idea is very simple: Partition the points  $u_i$  by a random hyperplane. Equivalently, choose  $v \in \mathbb{S}^{n-1}$  uniformly at random and define  $x_i := \text{sgn}(\langle u_i, v \rangle)$ , where  $\text{sgn}(t) = +1$  if  $t \geq 0$  and  $\text{sgn}(t) = -1$  if  $t < 0$ .

---

<sup>11</sup>Observe also that the SDP is feasible (take any set of unit vectors and the resulting matrix of scalar products) and bounded: since the  $u_i$ 's are unit vectors, we have  $|y_{ij}| = |\langle u_i, u_j \rangle| \leq 1$  for all  $i, j$ , hence  $|\sum_{i < j} a_{ij} y_{ij}| \leq |E| < \infty$ .

Thus, the approximation algorithm for MAXCUT consists of the following steps:

1. Given a graph  $G$ , compute an optimal solution  $Y$  to the SDP (2.11).
2. Compute a Gram decomposition  $Y = U^T U$ . The column vectors  $u_1, \dots, u_n \in \mathbb{S}^{n-1}$  form an optimal solution for (2.10). As mentioned above,  $\sum_{i < j} a_{ij} \frac{1 - \langle u_i, u_j \rangle}{2}$  is at least as large as the maximum size of a cut in  $G$ .
3. Use “random rounding” by a random vector  $v \in \mathbb{S}^{n-1}$ , i.e., define  $x_i := \text{sgn}(\langle u_i, v \rangle)$ . It will follow from Lemma 2.23 below that the resulting cut has size at least 0.878 times  $\sum_{i < j} a_{ij} \frac{1 - \langle u_i, u_j \rangle}{2}$ , hence at least 0.878 times the size of a maximum cut.

In order to make the description of the algorithm absolutely complete and precise, one would also need to specify how to choose a uniformly random unit vector  $v \in \mathbb{S}^{n-1}$  *algorithmically*. Since the entries of  $v$  will typically be irrational, we can again only do this approximately (to arbitrary precision. One of the simplest ways may be to choose a vector  $w \in \mathbb{R}^n$  according to the standard normal distribution (this is simple (again up to arbitrary precision) since we can choose the entries of  $w$  as independent 1-dimensional standard normal random variables) and then set  $v = \frac{w}{\|w\|}$ . In fact, one could skip the renormalization and work directly with  $w$ , since all we will need is that the distribution from which we pick the vector is rotationally symmetric.

**Lemma 2.23.** *Let  $u \neq u' \in \mathbb{S}^{n-1}$ , and let  $v \in \mathbb{R}^n$  be a random vector chosen to a fixed rotationally symmetric (with respect to the origin) probability distribution (for instance, the uniform distribution on the unit sphere, or the standard normal distribution).<sup>12</sup>*

*Set  $x := \text{sgn}(\langle v, u \rangle)$  and  $x' := \text{sgn}(\langle v, u' \rangle)$  Then*

$$\mathbf{E} \left[ \frac{1 - xx'}{2} \right] = \Pr[xx' = -1] = \frac{\arccos \langle u, u' \rangle}{\pi} \geq \alpha \frac{1 - \langle u, u' \rangle}{2},$$

where  $\alpha := \inf_{-1 \leq s < 1} \frac{2 \arccos(s)}{\pi(1-s)} > 0.87856$ .

*Proof.* The equality  $\mathbf{E} \left[ \frac{1 - xx'}{2} \right] = \Pr[xx' = -1]$  is just the definition of the expectation of a 0/1-valued random variable. The inequality  $\frac{\arccos \langle u, u' \rangle}{\pi} \geq \alpha \frac{1 - \langle u, u' \rangle}{2}$  follows directly from the definition of  $\alpha$ , and the estimate for  $\alpha$  is the subject of Exercise 16.

It remains to prove the middle equation  $\Pr[xx' = -1] = \frac{\arccos \langle u, u' \rangle}{\pi}$ . Consider first the case that  $n = 2$ , i.e., that  $u$  and  $u'$  are unit vectors on the unit circle in

---

<sup>12</sup>Rotational symmetry means that if  $A \subseteq \mathbb{R}^d$  is a measurable set, and if  $\rho$  is a rotation of  $\mathbb{R}^d$  (an orthogonal linear transformation with determinant 1), then  $\Pr[v \in A] = \Pr[v \in \rho A]$ . We also assume that the origin has mass 0, i.e.,  $\Pr[v = \mathbf{0}] = 0$ .

the plane. We have  $xx' = -1$  iff the line orthogonal to  $v$  separates  $u$  and  $u'$  iff  $v$  lies in the double wedge formed by the two lines orthogonal to  $u$  and  $u'$ , respectively, see Figure 2.4. If  $\beta = \arctan(\langle u, u' \rangle)$  is the angle between  $u$  and  $u'$ , then this double wedge covers an angle of  $2\beta$ . Since the distribution of  $v$  is rotationally symmetric, the probability that  $v$  falls into this double wedge equals  $\frac{2\beta}{2\pi} = \beta/\pi$ , as desired.

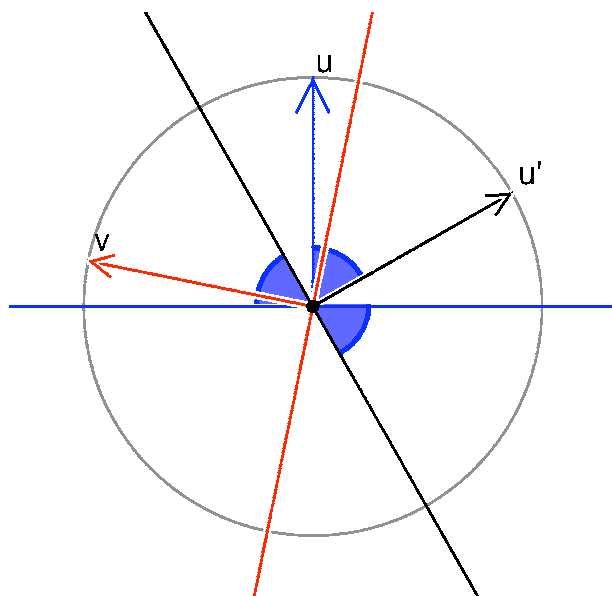


Figure 2.4: The probability that a random line separates two vectors equals twice the angle between them over  $2\pi$ .

In the general case  $u, u' \in \mathbb{S}^{n-1}$ , let  $F \subset \mathbb{R}^n$  be the 2-dimensional plane spanned by  $u$  and  $u'$ . Let  $\bar{v}$  be the orthogonal projection of  $v$  onto the plane  $F$ . Note that  $\langle u, \bar{v} \rangle = \langle u, v \rangle$  and  $\langle u', \bar{v} \rangle = \langle u', v \rangle$ . Moreover, the distribution of  $\bar{v}$  is again rotationally symmetric. Thus, the general statement follows from the 2-dimensional case.  $\square$

## 2.6 Lower Bounds For Euclidean Embeddings

When defining the notion of distortion we remarked that, for example, the graphs in Figure 2.3 are not isometrically embeddable into Euclidean of any dimension, and we left the proof of this statement as an exercise. In this section, we will present a method to prove lower bounds for the minimal distortion necessary to embed a given metric space into Euclidean space. We will prove the following:

**Theorem 2.24.** Let  $C^m := (\{0, 1\}^m, \rho_H)$  be the  $m$ -dimensional Hamming cube.<sup>13</sup> Then any embedding  $f : C_m \rightarrow (\mathbb{R}^d, \|\cdot\|_2)$  into Euclidean space has distortion at least  $\sqrt{m}$ .

Thus, the theorem says that the “natural embedding”, where we simply consider  $\{0, 1\}^m$  as a subset of  $\mathbb{R}^m$ , has the minimal distortion: observe that for this embedding the lengths of the “edges” are preserved: we have  $\|x - y\|_2 = 1$  if  $x$  and  $y$  are adjacent in the Hamming cube, i.e., if  $\rho_H(x, y) = 1$ ; on the other hand, the distances between all non-adjacent pairs are contracted: if  $\rho_H(x, y) = k$  then  $\|x - y\|_2 = \sqrt{k}$ ; the contraction is strongest for the “main diagonals”, i.e., for pairs of points at maximal Hamming distance. Intuitively, for any embedding, we either into Euclidean space, either we have to expand the edges or to contract the main diagonals of the Hamming cube.

To turn this kind of intuition into a rigorous proof, the general proof strategy is the following. Let  $(V, \rho)$  be a finite metric space. For a collection  $E \subseteq \binom{V}{2}$ , consider the following “average” of squared distances between point pairs in  $E$ :

$$\text{ave}_2(\rho, E) := \sqrt{\frac{\sum_{\{u,v\} \in E} \rho(u, v)^2}{|E|}}.$$

As we will see below, for embeddings into Euclidean spaces, squared distances are technically very convenient. Now, if  $E, F \subseteq \binom{V}{2}$  are two collections of pairs, we consider the ratio

$$R_{F,E}(\rho) := \frac{\text{ave}_2(\rho, F)}{\text{ave}_2(\rho, E)}.$$

The idea is that we want to capture, in a quantitative way, a certain kind of “trade-off”, as between the lengths of the edges versus the diagonals in the case of the Hamming cube.

Any  $f : V \rightarrow \mathbb{R}^d$  gives induces a second metric on  $V$ , the Euclidean distances between the image points:

$$\sigma(u, v) := \|f(u) - f(v)\|_2,$$

and we can also consider the quantities  $\text{ave}_2(\sigma, E)$  and  $R_{F,E}(\sigma)$ .

**Lemma 2.25.** If  $f$  has distortion at most  $D$ , then

$$\frac{1}{D} R_{F,E}(\rho) \leq R_{F,E}(\sigma) \leq D \cdot R_{F,E}(\rho).$$

---

<sup>13</sup>Recall that the Hamming distance  $\rho_H(x, y) = |\{i : x_i \neq y_i\}|$  between two 0/1-strings of length  $m$  is the number of positions in which they differ.

*Proof.* By rescaling the map  $f$ , if necessary (the target space is a normed space), we may assume that  $\rho(x, y) \leq \sigma(x, y) = \|f(x) - f(y)\|_2 \leq D\rho(x, y)$  for all  $x, y \in V$ . Applying the first inequality for pairs in  $E$  and the second one for pairs in  $F$ , we see that  $\text{ave}_2(\sigma, E) \geq \text{ave}_2(\rho, E)$  and  $\text{ave}_2(\sigma, F) \leq D \cdot \text{ave}_2(\rho, F)$ . Thus, Thus,

$$R_{F,E}(\sigma) = \frac{\text{ave}_2(\sigma, F)}{\text{ave}_2(\sigma, E)} \leq D \cdot \frac{\text{ave}_2(\rho, F)}{\text{ave}_2(\rho, E)} = D \cdot R_{F,E}(\rho)$$

If we apply the inequalities defining distortion the other way around, we obtain  $\text{ave}_2(\sigma, E) \leq D \cdot \text{ave}_2(\rho, E)$  and  $\text{ave}_2(\sigma, F) \geq \text{ave}_2(\rho, F)$ , and hence

$$R_{F,E}(\sigma) = \frac{\text{ave}_2(\sigma, F)}{\text{ave}_2(\sigma, E)} \geq \frac{\text{ave}_2(\rho, F)}{D \cdot \text{ave}_2(\rho, E)} = \frac{1}{D} R_{F,E}(\rho)$$

□

In order to apply this lemma to the Hamming cube, we will need the following simple statement:

**Lemma 2.26** (Short Diagonals Lemma). *Let  $p_1, p_2, p_3, p_4 \in \mathbb{R}^d$ . Then*

$$\|p_1 - p_3\|_2^2 + \|p_2 - p_4\|_2^2 \leq \|p_1 - p_2\|_2^2 + \|p_2 - p_3\|_2^2 + \|p_3 - p_4\|_2^2 + \|p_4 - p_1\|_2^2.$$

Thus, for any embedding of the vertices of a square into Euclidean space, the sum squared lengths of the diagonals is at most the sum of the squared edge lengths. Note first that the inequality is easy to verify in dimension  $d = 1$ : If  $x_1, x_2, x_3, x_4 \in \mathbb{R}^1$  are four real numbers then the difference between the right-hand side and the left-hand side of the inequality is

$$\begin{aligned} (x_1 - x_2)^2 + (x_2 - x_3)^2 + (x_3 - x_4)^2 + (x_4 - x_1)^2 - (x_1 - x_3)^2 - (x_2 - x_4)^2 \\ = (x_1 - x_2 + x_3 - x_4)^2 \geq 0, \end{aligned}$$

by collecting terms. Moreover, the squared Euclidean norm of a vector in  $\mathbb{R}^d$  is just the sum of squares of its coordinates. Thus, the lemma follows from the 1-dimensional case by considering each coordinate separately. We leave the details as an exercise. (Hint: First try the three-dimensional case and write the coordinates of each point as  $p_i = (x_i, y_i, z_i)$ . Then the inequality we want to prove boils down to three 1-dimensional inequalities, one involving the  $x$ 's, one involving the  $y$ 's, and one involving the  $z$ 's. The higher-dimensional case is exactly the same, only that it becomes more cumbersome to write out the coordinates. The details are left as an exercise.)

*Proof.* Proof of Theorem 2.24 Let  $V = \{0, 1\}^m$ , let  $E = \{\{u, v\} \in \binom{V}{2} : \rho_H(u, v) = 1\}$  be the edge set of the Hamming cube (pairs of vertices that disagree in 1 coordinate), and let  $F := \{\{u, \bar{u}\} \in \binom{V}{2} : \rho_H(u, \bar{u}) = m\}$  be the set of "main

diagonals" of the Hamming cube (pairs of vertices that disagree in all coordinates). Observe that  $|E| = m2^{m-1}$  (every vertex is incident to precisely  $m$  edges) and  $|F| = 2^{m-1}$  (every vertex is incident to precisely one main diagonal). Moreover,  $\text{ave}_2(\rho_H, E) = 1$  and  $\text{ave}_2(\rho_H, F) = m$ , hence  $R_{F,E}(\rho_H) = m$ .

Assume now that there is a map  $f : C^m \rightarrow (\mathbb{R}^d, \|\cdot\|_2)$  of distortion at most  $D$ . We want to show that  $D \geq \sqrt{m}$ . Note that any map has distortion at least 1, so the case  $m = 1$  is trivial and we will assume  $m \geq 2$  from now on. Let  $\sigma$  be the metric on  $V$  induced by  $f$ ,  $\sigma(u, v) = \|f(u) - f(v)\|_2$ .

**Claim.**  $\sum_{\{u,v\} \in E} \sigma(u, v)^2 \geq \sum_{\{u,\bar{u}\} \in F} \sigma(u, \bar{u})^2$ .

Before we prove this claim, observe that it implies

$$\text{ave}_2(\sigma, F)^2 = \frac{\sum_{\{u,\bar{u}\} \in F} \sigma(u, \bar{u})^2}{2^{m-1}} \leq m \cdot \frac{\sum_{\{u,v\} \in E} \sigma(u, v)^2}{m2^{m-1}} = m \cdot \text{ave}_2(\sigma, E),$$

hence

$$R_{E,F}(\sigma) = \frac{\text{ave}_2(\sigma, F)}{\text{ave}_2(\sigma, E)} \leq \sqrt{m}.$$

Since  $R_{F,E}(\rho_H) = m$ , it follows that  $R_{F,E}(\sigma) \leq \frac{1}{\sqrt{m}} R_{F,E}(\rho)$ , so it follows from Lemma 2.25 that the map  $f$  has distortion at least  $\sqrt{m}$ , which is what we want to prove. So it remains to prove the claim, which we will do by induction on  $m$ .

The base case  $m = 2$  is precisely the Short Diagonals Lemma. For the induction step, we split the vertex set  $V = \{0, 1\}^m$  into two subsets  $V_0 = \{u \in V : u_m = 0\}$  and  $V_1 = \{u \in V : u_m = 1\}$ . Note that these two vertex sets induce two disjoint subgraphs  $C_i^{m-1}$ ,  $i = 0, 1$  of  $C^m$ , each of which is isomorphic to the  $(m-1)$ -dimensional Hamming cube. Namely, the edge set of  $C_i^{m-1}$  is  $E_i = \{\{u, v\} \in E : u, v \in V_i\} = \{\{u, v\} \subseteq V_i : \rho_H(u, v) = 1\}$ . Within each  $C_i^{m-1}$ , we have the set of its  $((m-1)$ -dimensional) main diagonals  $F_i = \{\{u, v\} \in \binom{V_i}{2} : \rho_H(u, v) = m-1\}$ ,  $i = 0, 1$ . Moreover, we denote the set of the remaining edges of  $C^m$  (which connect  $V_0$  and  $V_1$ ) by  $E_{01} := E \setminus (E_0 \cup E_1)$ , the set of remaining edges of  $C^m$ .

Moreover, consider a main diagonal  $\{u, \bar{u}\} \in F$  in  $C^m$ . Without loss of generality, suppose that  $u \in V_0$  and  $\bar{u} \in V_1$ . The vertex  $u$  has a unique "upstairs" neighbor  $v \in V_1$ ,  $\{u, v\} \in E_{01} \subseteq E$ , and the vertex  $\bar{u}$  has a unique "downstairs" neighbor  $\bar{v} \in V_0$ ,  $\{\bar{u}, \bar{v}\} \in E_{01} \subseteq E$ , and  $\{v, \bar{v}\} \in F$  is another main diagonal in  $C^m$ . Moreover, the  $\{u, \bar{v}\} \in F_0$  and  $\{\bar{u}, v\} \in F_1$  form main diagonals in the  $(m-1)$ -dimensional subcubes, see Figure 2.5. Moreover, by the Short Diagonals Lemma,

$$\sigma(u, \bar{u})^2 + \sigma(v, \bar{v})^2 \leq \sigma(u, v)^2 + \sigma(\bar{u}, \bar{v})^2 + \sigma(u, \bar{v})^2 + \sigma(v, \bar{u})^2.$$

When summing up over all pairs  $\{u, \bar{u}\}$  and  $\{v, \bar{v}\}$  of  $m$ -dimensional main

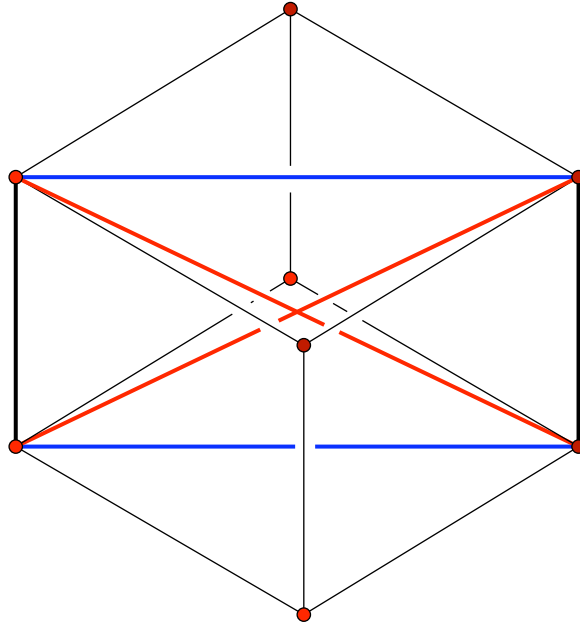


Figure 2.5: Two main diagonals with adjacent endpoints form the diagonals of a square whose sides are two edges of the cube and two facet diagonals.

diagonals with  $\rho_H(u, v) = \rho_H(\bar{u}, \bar{v}) = 1$ , we consider each edge in  $F$ ,  $E_{01}$ ,  $F_0$ , and  $F_1$ , respectively, exactly once. Thus,

$$\sum_{\{u, \bar{u}\} \in F} \sigma(u, \bar{u})^2 \leq \sum_{\{u, v\} \in E_{01}} \sigma(u, v)^2 + \sum_{\{u, \bar{v}\} \in F_0} \sigma(u, \bar{v})^2 + \sum_{\{\bar{u}, v\} \in F_1} \sigma(\bar{u}, v)^2$$

Moreover, by the induction hypothesis, the last two sums are bounded from above by  $\sum_{\{u, v\} \in E_0} \sigma(u, v)^2$  and  $\sum_{\{u, v\} \in E_1} \sigma(u, v)^2$ , respectively. Since  $E = E_{01} \cup E_0 \cup E_1$  is a partition, we conclude that

$$\sum_{\{u, \bar{u}\} \in F} \sigma(u, \bar{u})^2 \leq \sum_{\{u, v\} \in E} \sigma(u, v)^2,$$

as desired. This proves the claim and hence the theorem.  $\square$

**Remark 2.27.** As we have seen, Hamming cubes are examples of metric spaces with  $n = 2^m$  elements that cannot be embedded into any Euclidean space with distortion less than  $\sqrt{m} = \sqrt{\log_2(n)}$ . The same proof method of quadratic averages can be used to show that there are  $n$ -element metric space cannot be embedded into any Euclidean space with distortion less than  $\Omega(\log(n))$ . Examples for such metric spaces are *constant-degree expander graphs* with the shortest path metric, see [7, Section 15.5]

## 2.7 Upper Bounds For Euclidean Embeddings

As remarked at the end of the last section, there are  $n$ -point metric spaces that require a distortion of  $\Omega(\log n)$  for any embedding into Euclidean space. The goal of this section is to prove a matching upper bound, due to Bourgain.

**Theorem 2.28.** *Every  $n$ -point metric space  $(V, \rho)$  can be embedded into some Euclidean space  $(\mathbb{R}^d, \|\cdot\|_2)$  with distortion at most  $O(\log n)$ .*

The proof we will present in detail yields a target dimension of  $d = 2^n$ . With a little bit of more work, this can be pushed down to  $d = O(\log^2 n)$ . Moreover, with minor modifications the same proof also works for embeddings into  $(\mathbb{R}^d, \|\cdot\|_p)$  for any  $p$ -norm,  $1 \leq p \leq \infty$ , and also yields a randomized algorithm to construct such embeddings. We will say more about this along the way, while leaving some of the details as exercises.

We first introduce the notion of a *line pseudometric* that will play a crucial role in the proof. Generally, a *pseudometric* on a set  $V$  is a map  $\nu : V \times V \rightarrow \mathbb{R}$  that satisfies all the requirements of a metric (nonnegativity, symmetry, and the triangle inequality), except that we allow the possibility that  $\nu(x, y) = 0$  for  $x \neq y$ . A *line pseudometric* is a pseudometric of the form  $\nu(x, y) = |\varphi(x) - \varphi(y)|$  for some map  $\varphi : V \rightarrow \mathbb{R}$ .

In fact, we will be working with line pseudometrics of a special kind that we already encountered when showing that every  $n$ -point metric space can be isometrically embedded into  $\mathbb{R}^{2^n}$  with the  $\|\cdot\|_\infty$  norm (Exercise 12): For a subset  $A \subseteq V$  of the metric space, define  $\varphi_A(x) := \rho(x, A) := \min_{a \in A} \rho(x, a)$  and  $\nu_A(x, y) := |\rho(x, A) - \rho(y, A)|$ . These special line pseudometrics have the additional property that they are dominated by  $\rho$  in the sense that<sup>14</sup>  $\nu_A \leq \rho$ .

We cannot hope to encode too much information about a metric  $\rho$  on  $V$  by means of a single pseudometric (which we can think of as a 1-dimensional “projection”), but the following lemma tells us that if we can approximate the  $\rho$  given metric by a convex combination of line pseudometrics each of which is dominated by  $\rho$ , then we can use the projections as coordinates for an embedding into Euclidean space.

**Lemma 2.29.** *Let  $(V, \rho)$  be a finite metric space, and let  $\nu_1, \dots, \nu_d$  be line pseudometrics on  $V$  such that  $\nu_i \leq \rho$  for all  $i$  and that*

$$\sum_{i=1}^d \alpha_i \nu_i \geq \frac{1}{D} \rho.$$

*for some coefficients  $\alpha_i \geq 0$  with  $\sum_i \alpha_i = 1$ . Then there is an embedding  $f : (V, \rho) \rightarrow (\mathbb{R}^d, \|\cdot\|_2)$  with distortion at most  $D$ .*

<sup>14</sup>That is,  $\nu_A(x, y) \leq \rho(x, y)$  for all  $x, y \in X$ . To see this, consider some  $a \in A$  with  $\rho(y, A) = \rho(y, a)$ . Then  $\rho(x, A) \leq \rho(x, a) \leq \rho(x, y) + \rho(y, a) = \rho(x, y) + \rho(y, A)$ , hence  $\rho(x, A) - \rho(y, A) \leq \rho(x, y)$ . The inequality  $\rho(y, A) - \rho(x, A) \leq \rho(x, y)$  follows symmetrically.

*Proof.* For  $1 \leq i \leq d$ , let  $\varphi_i : V \rightarrow \mathbb{R}$  be a map inducing the line pseudometric  $\nu_i$ . We define  $f : V \rightarrow \mathbb{R}^d$  by

$$f(v) = (\sqrt{\alpha_1}\varphi_1(v), \dots, \sqrt{\alpha_d}\varphi_d(v)).$$

Then, by assumption on the  $\alpha_i$  and the  $\nu_i$ ,

$$\|f(u) - f(v)\|_2^2 = \sum_{i=1}^d \alpha_i \nu_i(u, v)^2 \leq \underbrace{\left( \sum_{i=1}^d \alpha_i \right)}_{=1} \rho(u, v)$$

for all  $u, v \in V$ . Moreover, by the Cauchy-Schwarz inequality,

$$\begin{aligned} \frac{1}{D} \rho(u, v) &\leq \sum_{i=1}^d \alpha_i \nu_i(u, v) = \sum_{i=1}^d \sqrt{\alpha_i} \left( \sqrt{\alpha_i} \nu_i(u, v) \right) \\ &\leq \underbrace{\left( \sum_{i=1}^d \alpha_i \right)}_{=1}^{1/2} \left( \sum_{i=1}^d \alpha_i \nu_i(u, v)^2 \right)^{1/2} = \|f(u) - f(v)\|_2 \end{aligned}$$

□

**Remark 2.30.** Essentially the same proof (with Hölder's Inequality instead of Cauchy-Schwarz) shows that under the assumptions of the lemma, for any  $1 \leq p \leq \infty$  there exists an embedding of  $(V, \rho)$  into  $(\mathbb{R}^d, \|\cdot\|_p)$  with distortion at most  $D$  (Exercise 21).

As remarked above, we will use line pseudometrics of the form  $\nu_A(u, v) = |\rho(u, A) - \rho(v, A)|$  for subsets  $A \subseteq V$ . Here is the main technical lemma:

**Lemma 2.31.** *Let  $u, v \in V, u \neq v$ . Then there exist real numbers  $\Delta_1, \Delta_2, \dots, \Delta_q \geq 0$  with  $\sum_i \Delta_i = \frac{1}{4}\rho(u, v)$ , where  $q = \lfloor \log_2 n \rfloor + 1$ , such that the following holds for all  $1 \leq j \leq q$ :*

*Let  $A_j$  be a random subset of  $V$ , where we include each point of  $V$  independently with probability  $2^{-j}$ . Then*

$$\Pr\left[\underbrace{|\rho(u, A_j) - \rho(v, A_j)|}_{\nu_{A_j}(u, v)} \geq \Delta_j\right] \geq \frac{1}{12}.$$

*Consequently, the expectation of the nonnegative random variable  $\nu_{A_j}(u, v)$  satisfies  $\mathbf{E}[\nu_{A_j}(u, v)] \geq \frac{1}{12}\Delta_j$ .*

Before we prove this lemma, let us first see how it allows us to choose the convex combination of line pseudometrics that yields the desired low-distortion embedding.

*Proof that Lemma 2.31 implies Theorem 2.28.* For each subset  $A \subseteq V$ , let  $\nu_A(u, v) = |\rho(u, A) - \rho(v, A)|$  be the corresponding line pseudometric. We want to apply Lemma 2.31. For  $1 \leq j \leq q$ , let  $\pi_j(A) := \Pr[A_j = A]$ , where  $A_j$  is the random subset with point probability  $2^{-j}$  as in the lemma. Then by the lemma, for every pair of distinct points  $u, v \in V$ ,

$$\sum_{A \subseteq V} \pi_j(A) \cdot \nu_A(u, v) = \mathbf{E}[\nu_{A_j}(u, v)] \geq \frac{1}{12} \Delta_j.$$

Summing over all  $1 \leq j \leq q$  and exchanging the order of summation, we conclude

$$\sum_{A \subseteq V} \left( \sum_{j=1}^q \pi_j(A) \right) \cdot \nu_A(u, v) \geq \frac{1}{12} \sum_{j=1}^q \Delta_j = \frac{1}{48} \rho(u, v).$$

Dividing by  $q$ , we obtain

$$\sum_{A \subseteq V} \underbrace{\left( \frac{1}{q} \sum_{j=1}^q \pi_j(A) \right)}_{=: \alpha_A} \nu_A(u, v) \geq \frac{1}{48q}$$

for all  $u, v \in V$ . Moreover, for each  $j$ , we have  $\sum_{A \subseteq V} \pi_j(A) = 1$ , hence  $\sum_{A \subseteq V} \alpha_A = 1$ . Thus, by Lemma `lem:convex-comb-line-pseudometrics`, there is an embedding  $f : (V, \rho) \rightarrow (\mathbb{R}^d, \|\cdot\|_2)$ ,  $d = 2^n$ , with distortion at most  $48q = O(\log n)$ .  $\square$

**Remark 2.32.** The proof as it stands yields an embedding dimension of  $d = 2^n$ . For embeddings into Euclidean spaces, we can in a second step apply the Johnson-Lindenstrauss Flattening Theorem and project the  $n$ -point set  $f(V) \subset \mathbb{R}^{2^n}$  obtained in the first step into  $\mathbb{R}^{O(\log n)}$ , with an additional distortion factor of 2, say. For  $p$ -norms with  $p \neq 2$ , no Flattening Theorem is available, but the target dimension can still be brought down to  $O(\log^2 n)$ : For each  $1 \leq j \leq q$ , consider  $m = O(\log n)$  independent random subsets  $A_j^1, \dots, A_j^m$  with point probability  $2^{-j}$  each. Using suitable Chernov-type estimates (recall Lemma 2.16) and the preceding lemmas, one can show that with high probability, the map  $f : V \rightarrow \mathbb{R}^{mq}$  with coordinate  $f_{ij}(u) := \rho(u, A_j^i)$  has distortion at most  $O(\log n)$  with respect to any given, fixed  $p$ -norm  $\|\cdot\|_p$ ,  $1 \leq p \leq \infty$ . Details are left as Exercise 22.

Since we can check the distortion of a given map  $(V, \rho) \rightarrow (\mathbb{R}^d, \|\cdot\|_p)$  efficiently (in the most simple-minded way, by computing all the pairwise distances in the target space), we immediately obtain a randomized algorithm that computes an embedding with  $O(\log n)$ -distortion in expected polynomial time (choose the sets  $A_j^i$  as above, compute the distortion of the resulting map, and if it is too large, repeat.)

**Remark 2.33** (Euclidean Embeddings and Semidefinite Programming). One can show that the minimum distortion  $D$  necessary for embedding a given metric space into Euclidean space can be determined efficiently using semidefinite programming (Exercise 24). We mention that using this method, the target dimension for the embedding cannot be described in advance.

Summarizing, we obtain the following:

**Theorem 2.34.** *If  $(V, \rho)$  is an  $n$ -point metric space and if  $1 \leq p \leq \infty$ , then there is a mapping  $f : (V, \rho) \rightarrow (\mathbb{R}^d, \|\cdot\|_p)$  of distortion at most  $O(\log n)$ , where  $d = O(\log^2 n)$ ; for  $p = 2$ , the target dimension can be reduced to  $O(\log n)$ . Moreover, there is a randomized algorithm that computes the mapping in expected polynomial time.*

*Proof of Lemma 2.31.* Fix  $u, v \in V$ ,  $u \neq v$ . Set  $r_q := \frac{1}{4}\rho(u, v)$ . For  $0 \leq j \leq q - 1$ , define  $\tilde{r}_j$  to be the smallest radius such that both  $|B_{\tilde{r}_j}(u)| \geq 2^j$  and  $|B_{\tilde{r}_j}(v)| \geq 2^j$ , where  $B_r(x) = \{y \in V : \rho(x, y) \leq r\}$ . Set  $r_j = \min\{r_q, \tilde{r}_j\}$ . We will show that the lemma is true with  $\Delta_j := r_j - r_{j-1}$ .

Fix  $j \in \{1, \dots, q\}$ . We may assume that  $r_{j-1} = \tilde{r}_{j-1} < r_q$ , since otherwise  $\Delta_j = 0$  and the conclusion of the lemma holds trivially. Consequently, both closed balls  $B_{r_{j-1}}(u)$  and  $B_{r_{j-1}}(v)$  contain at least  $2^{j-1}$  points. Let  $A_j$  be a random subset of  $V$  with point probability  $2^{-j}$ . By definition of  $r_j$ , at least one of the open balls  $B_{r_j}^\circ(u) = \{x \in V : \rho(x, u) < r_j\}$  and  $B_{r_j}^\circ(v)$  contains less than  $2^j$  points, say  $|B_{r_j}^\circ(u)| < 2^j$  (this also holds for  $j = q$ , since  $|V| \leq 2^q$ ). Call  $A_j$  *good* if  $A_j \cap B_{r_{j-1}}(v) \neq \emptyset$  and  $A_j \cap B_{r_j}^\circ(u) = \emptyset$ . The former ball has cardinality at least  $2^{j-1}$  and the latter at most  $2^j$ , and the two balls are disjoint. Therefore,

$$\Pr[A_j \cap B_{r_{j-1}}(v) \neq \emptyset] = 1 - (1 - 2^{-j})^{2^{j-1}} \geq 1 - e^{-2^{-j}2^{j-1}} \geq 1 - e^{-1/2} > 1/3.$$

On the other hand,

$$\Pr[A_j \cap B_{r_j}^\circ(u) = \emptyset] \geq (1 - 2^{-j})^{2^j} \geq 1/4,$$

since  $0 \leq 2^{-j} \leq 1/2$ . The two events are independent, so we may multiply their probabilities and conclude that with probability at least  $1/12$ ,  $A_j$  is good. But then  $\rho(u, A_j) \geq r_j$  and  $\rho(v, A_j) \leq r_{j-1}$ , so  $\nu_{A_j}(u, v) \geq \Delta_j$ .  $\square$

## 2.8 Exercises

**Exercise 10.** *Consider the two 4-point metric spaces given by the shortest path metrics on the two graphs in Figure 2.3. Show that neither of these metric spaces can be isometrically embedded into  $\ell_2^d$ , for any  $d$ .*

**Exercise 11.** *Let  $f : X \rightarrow Y$  be a bijective map between vector spaces. Show that the distortion of  $f$  equals  $\|f\|_{Lip} \cdot \|f^{-1}\|_{Lip}$ .*

**Exercise 12.** Show that for every  $d \geq 1$ , there is an isometry  $f : \ell_1^d \rightarrow \ell_\infty^{2^d}$ . (An isometry is a distance preserving or distortion 1 map, i.e., in our case we require that  $\|f(x) - f(y)\|_\infty = \|x - y\|_1$  hold for all  $x, y \in \mathbb{R}^d$ .)

**Exercise 13.** Suppose you wish to design an algorithm that solves the following problem: Given a finite set  $X \subseteq \ell_1^d$ ,  $|X| = n$ , compute the diameter  $\text{diam}(X) := \max_{x, y \in X} \|x - y\|_1$ . What is the runtime of the “naive” algorithm that just computes pairwise distances? Show, using Exercise 12, that there is an algorithm that computes the diameter of a set of  $n$  points in  $\ell_1^d$  in time  $O(d2^d n)$ . (If  $d$  is fixed and  $n$  is large, this improves upon the naive algorithm.)

**Exercise 14.** Let  $(X, \rho)$  be an arbitrary metric space,  $|X| = n$ . Show that there is an isometry  $f : X \rightarrow \ell_\infty^n$ .

**Exercise 15.** Let  $A = [a_{ij}]$  be a symmetric  $(n \times n)$ -matrix.  $A$  is called positive semidefinite, written  $A \succcurlyeq 0$ , if  $x^T A x \geq 0$  holds for all  $x \in \mathbb{R}^n$  (where  $\cdot^T$  denotes the transpose).

(a) Show that the set  $SDF_n$  of positive semidefinite symmetric  $(n \times n)$ -matrices is a convex cone, i.e., if  $A, B \in SDF_n$  and if  $\alpha, \beta \in \mathbb{R}_{\geq 0}$ , then  $\alpha A + \beta B \in SDF_n$ .

(b) Show that the following statements are equivalent:

(i)  $A \succcurlyeq 0$ .

(ii) All eigenvectors of  $A$  are nonnegative.

(iii) For some dimension  $d$ , there exist vectors  $u_1, \dots, u_n \in \mathbb{R}^d$  such that  $A$  is the Gram matrix of the  $u_i$ 's, in the sense that  $a_{ij} = \langle u_i, u_j \rangle$  for  $1 \leq i, j \leq n$ . In other words,  $A = U^T U$  for some  $U \in \mathbb{R}^{d \times n}$ . (A bonus question: What is the minimal  $d$  (as a well-known function of  $A$ ) such that  $A$  can be written as such a Gram matrix?)

(iv)  $A$  can be written as a nonnegative linear combination of matrices of the form  $vv^T$ ,  $v \in \mathbb{R}^n$ .

**Exercise 16.** Define

$$\alpha := \inf_{-1 \leq s < 1} \frac{2 \arccos(s)}{\pi(1-s)} = \inf_{0 < t \leq \pi} \frac{2t}{\pi(1 - \cos(t))}.$$

Show that  $\alpha > 0.87856$ .

(To avoid ambiguities, we mean the function  $\arccos : [-1, 1] \rightarrow [0, \pi]$  that is the inverse of the cosine function  $\cos : [0, \pi] \rightarrow [-1, 1]$ .)

**Exercise 17.** Show that the following inequality holds for all  $v_1, v_2, v_3, v_4 \in \mathbb{R}^d$ :

$$\|v_1 - v_3\|_2^2 + \|v_2 - v_4\|_2^2 \leq \|v_1 - v_2\|_2^2 + \|v_2 - v_3\|_2^2 + \|v_3 - v_4\|_2^2 + \|v_4 - v_1\|_2^2.$$

(Hint: The above inequality can be expressed as a sum of 1-dimensional inequalities, one for each coordinate.)

The goal of the next three exercises is to show that the bound on the target dimension in the Johnson-Lindenstrauss Flattening Theorem cannot be improved much.

**Exercise 18.** (a) Let  $A = [a_{ij}]$  be a symmetric real  $(n \times n)$ -matrix such that  $a_{ii} = 1$  for all  $i$  and  $|a_{ij}| \leq 1/\sqrt{n}$  for  $i \neq j$ . Show that  $\text{rank } A > n/2$ . (Hint: Let  $\lambda_1, \dots, \lambda_r$  be the non-zero eigenvalues of  $A$ . Show, using traces of matrices, that  $\sum_{i=1}^r \lambda_i = n$  and  $\sum_{i=1}^r \lambda_i^2 < 2n$ , and apply the Cauchy-Schwarz Inequality in a clever way.)

(b) Suppose that  $A$  satisfies the assumptions from Part (a), except that  $A$  is not necessarily symmetric. Show that we can still conclude  $\text{rank}(A) > n/4$ . (Hint: How do you symmetrize a matrix? By which factor can its rank increase at most in the process?)

**Exercise 19.** Let  $B = [b_{ij}]$  be a real  $(n \times n)$ -matrix, and let  $f(x) \in \mathbb{R}[x]$  be a polynomial (in one variable) of degree  $k$ . Define a matrix  $C = [c_{ij}]$  by applying  $f$  to each entry of  $B$  separately, i.e.,  $c_{ij} = f(b_{ij})$  for  $1 \leq i, j \leq n$ . Show that

$$\text{rank}(C) \leq \binom{k+r}{r} \leq \left( \frac{e(k+r)}{k} \right)^k.$$

(Hint: Suppose without loss of generality that the first  $r$  rows of  $B$  linearly span all other rows. For integers  $0 \leq k_1, \dots, k_r$  with  $\sum_{i=1}^r k_i \leq k$ , define a vector  $v^{(k_1, \dots, k_r)} \in \mathbb{R}^n$  by  $v_j^{(k_1, \dots, k_r)} := \prod_{i=1}^r b_{ij}^{k_i}$  and show that the span of these vectors contains all rows of  $C$ .)

**Exercise 20.** (a) Let  $B = [b_{ij}]$  be a real  $(n \times n)$ -matrix such that  $b_{ii} = 1$  for all  $i$  and  $|b_{ij}| \leq \varepsilon$  for  $i \neq j$ , where  $1/\sqrt{n} \leq \varepsilon \leq 1/2$ . Show that

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

(Hint: Apply Exercises 18 (b) and 19. To what power do you have to raise  $\varepsilon$  to make it smaller than  $1/\sqrt{n}$ ?)

(b) Consider the set  $X = \{0, e_1, \dots, e_n\} \subset \mathbf{R}^n$  (where the  $e_i$ 's are the vectors of the standard orthonormal basis). Suppose that this set of points (with their Euclidean distances) can be mapped with distortion at most  $(1 + \varepsilon)$  into  $\ell_2^k$  (i.e., into  $\mathbb{R}^k$  with Euclidean distances). Show that then there exist  $v_1, \dots, v_n \in \mathbb{R}^k$  that are "almost orthogonal" unit vectors, i.e.,  $\|v_i\|_2 = 1$  for all  $i$  and  $|\langle v_i, v_j \rangle| \leq 100\varepsilon$  (the constant 100 could be improved).

(c) Assuming that there is a low-distortion map as in Part (b) and  $\frac{1}{100\sqrt{n}} \leq \varepsilon \leq 1/2$ , show that

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

**Exercise 21.** Extend Lemma 2.29 to arbitrary  $p$ -norms  $\|\cdot\|_p$ ,  $1 \leq p \leq \infty$ . That is, show that if  $(V, \rho)$  is a finite metric space and if  $\nu_1, \dots, \nu_d$  are line pseudometrics on  $V$  with  $\nu_i \leq \rho$  for all  $i$  and  $\sum_{i=1}^d \alpha_i \nu_i \geq \frac{1}{D} \rho$  for some  $\alpha_i \geq 0$  with  $\sum_i \alpha_i = 1$  and some  $D > 0$ , then  $(V, \rho)$  can be embedded into  $(\mathbb{R}^d, \|\cdot\|_p)$  with distortion at most  $D$ .

**Exercise 22.** (a) Let  $E_1, \dots, E_m$  be independent events in some probability space, each of them occurring with probability at least  $1/12$ . Show that the probability that at most  $\frac{m}{24}$  of the events occur at the same time is at most  $e^{-cm}$  for some suitable constant  $c > 0$ . Use Chernov-type bounds or direct estimates of binomial coefficients.

(b) Adapt the proof of Theorem 2.28 to reduce the target dimension to  $O(\log^2 n)$ : For  $q = \lfloor \log_2 n \rfloor + 1$  is as in Lemma 2.31 and  $1 \leq j \leq q$ , pick sets  $A_j^1, \dots, A_j^m \subseteq V$  independently at random, where each  $v \in V$  is included into  $A_j^i$  independently with probability  $2^{-j}$ . Using Part (a), show that if  $m = C \log n$  for a sufficiently large constant  $C > 0$ , then with high probability,

$$\sum_{i=1}^m \sum_{j=1}^q \frac{1}{mq} \nu_{A_j^i}(u, v) \geq \frac{1}{24q}$$

for all  $u, v \in V$ . Conclude, using Lemma 2.29 and Remark 2.30 that with high probability, the map  $f : V \rightarrow \mathbb{R}^{mq}$  with coordinates  $f(u)_{ij} = \rho(u, A_j^i)$  has distortion at most  $O(\log n)$ .

**Exercise 23.** (a) A cut pseudometric  $\tau$  on a set  $V$  is a pseudometric induced by a map  $\psi : V \rightarrow \{0, 1\}$ , i.e.,  $\tau(u, v) = |\psi(u) - \psi(v)|$  for all  $u, v \in V$ . Show that every line pseudometric  $\nu$  on an  $n$ -element set  $V$  is a nonnegative linear combination of at most  $n - 1$  cut pseudometrics, i.e.,  $\nu = \sum_{i=1}^{n-1} \alpha_i \tau_i$  for some  $\alpha_i \geq 0$  and cut pseudometrics  $\tau_i$ .

(b) Show that if  $V \subseteq \mathbb{R}^d$  and we consider distances in the 1-norm  $\|\cdot\|_1$ , then the resulting metric on  $V$  can be expressed as a nonnegative linear combination of at most  $d$  line pseudometrics.

**Exercise 24.** Let  $(V, \rho)$  be a finite metric space. Show that there is an embedding  $f : (V, \rho) \rightarrow (\mathbb{R}^d, \|\cdot\|_2)$  of distortion at most  $D$  for some target dimension  $d$  iff there is a positive definite matrix  $Q = [q_{uv}] \in \mathbb{R}^{V \times V}$  that satisfies

$$\rho(u, v)^2 \leq q_{uu} + q_{vv} - 2q_{uv} \leq D^2 \rho(u, v)^2$$

for all  $u, v \in V$ . Conclude that the minimum distortion necessary for embedding  $(V, \rho)$  into any Euclidean space can be determined in polynomial time by solving a semidefinite program.

Observe that with this method, the target dimension cannot be described.