# Dimensionality Reduction

Kent Quanrud

September 7, 2020

## 1 Large data sets and long vectors

Big data, big data, big data. What's so big about it? There are in fact two dimensions to be aware of. First, there a huge number of "pieces" of data being collected. For example, in the heavy hitters problem, we might have a piece of data for every search query made by a user. A second dimension that we have not yet confronted is the "size" or "width" of each piece of data. Here we will consider data where each piece of data is a high-dimensional array of real values; i.e., points in $\mathbb{R}^d$ for $d$ very large.

High-dimensional vectors arise rather easily. Every graph is associated with a square *adjacency matrix* whose dimensions are proportional to the number of vertices, $n$. Thus every row is an $n$-dimensional vector. The world wide web and social networks are by now extremely large graphs where the corresponding vectors have very high dimension. In text processing, text is sometimes represented as a "bag-of-words", where one counts the frequency of each word. This can be encoded as a feature vector whose dimensionality is proportionial to the size of the English language! (Plus typos.) To take this further, more aggressive algorithms use phrases - sequences of (say) 3 consecutive words - rather than words and run algorithms on "bag-of-phrases" vectors. These vectors have dimension proportional to the size of the English language, *cubed*! A recent technique from machine learning, called *autoencoders*, first trains a large model (such as a neural network) on some large collection of data. For each piece of data, the internal state of the model when labeling that data is ultimately a high dimensional vector. It has been observed that these high-dimensional vectors can have useful geometries; e.g., in the `word2vec` tool for word embeddings [5].

We note that in some of the examples above, the data vectors are typically sparse with few nonzero entries. Such vectors can be represented more compactly as an "adjacency list", where we list the index and the value of only the nonzero entries in the vector. The trouble arises when we start running computations over them. When we start combining these vectors in some linear algebraic procedure, the vectors rapidly become dense, and this is where we pay for the high dimensions.

Most operators with vectors take time proportional to the number of dimensions (in the worst and dense case). Certainly it would be desirable for the data to live in a much lower dimensional space. The goal in this discussion is to develop some techniques for transforming high-dimensional data into lower-dimensional data. We first note that for many applications, we do not necessarily require the exact coordinates of the vector. Given a set $P$ of points in a high-dimensional space $d$, we may only actually need the following:

1. For a given point $x \in P$, the (Euclidean) *length* of $x$, $\|x\| = \sqrt{\sum_i x_i^2}$.
2. For any two points $x, y \in P$, the Euclidean *distance* $\|x - y\|$ between them.

3. For any two points $x, y \in P$, the dot product $\langle x, y \rangle$ between $x$ and $y$.

Moreover, for many applications, approximations to the above quantities may suffice to produce approximation algorithms in the original objective.

We now introduce the main result of this article.

**Theorem 1** (Johnson and Lindenstrauss [1]). *Let $P \subseteq \mathbb{R}^d$ be a collection of $n$ points in $\mathbb{R}^d$, and let $k = O\big(\log(n)/\epsilon^2\big)$. Let $A \in \mathcal{N}^{k \times d}$ be a $k \times d$ randomized matrix where each coordinate is sampled as an independent Gaussian. Consider the randomly constructed linear map $f : \mathbb{R}^d \to \mathbb{R}^k$ defined by*

$$f(x) = \frac{1}{\sqrt{k}} Ax.$$

*Then with probability of error $\leq 1/\operatorname{poly}(n)$, we have*

$$(1 - \epsilon)\|x\| \leq \|f(x)\| \leq (1 + \epsilon)\|x\|$$

*for all $x \in P$, and*

$$(1 - \epsilon)\|x - y\| \leq \|f(x) - f(y)\| \leq (1 + \epsilon)\|x - y\|$$

*for all pairs $x, y \in \mathbb{R}^d$.*

This is a fairly remarkable theorem. Theorem 1 says that, for the sake of preserving distances, one can always reduce the dimension to about $\log(n)$, where $n$ is the number of input points. This bound is entirely independent of the input dimension. The input dimension could be as large as one could possibly imagine; the output dimension will always be a logarithmic function of the number of points. The construction, moreover, is *oblivious to the input*.

Perhaps even more remarkable is how *obvious* this mapping is after some acquantaince with Gaussian variables and their extremely convenient properties. The ideas underlying Theorem 1 lead to many other practical and simple (at least, to implement) algorithms, as we will see.

We note that the above guarantees also lead to approximations on pairwise dot-products; see Exercise 3.

We remark that the embedding $A$ given in Theorem 1 is not particularly compact, since it requires an independent Gaussian. This could be formiddably expension. Here one can instead replace the Gaussian entries with $\{-1, 0, 1\}$ entries generated by appropriate hash functions [2, 4, 6, 9]. The intuition is similar because $\{-1, 0, 1\}$-random variables behave similarly to Gaussian's in a certain technical sense. (They are both *sub-Gaussian*; see [10]). There is particular interest in ensuring that $A$ is *column-sparse*, since this determines the running time of applying $A$. An alternative approach uses *Hadamard matrices* to produce version of $A$ that can be applied extremely quickly [3].

An important application of dimensionality reduction is in accelerating numerical algorithms on large matrices. See for example [7, 8].

## 2 Gaussian random variables: an interface

Gaussian random variables are an *extremely convenient* class of random variables. To stress this point, rather than giving an explicit definition and proceeding with the mathematical analysis, we first outline (just a few of) the nice properties of Gaussian random variables, and put them to basic use. Later we will prove these properties, mostly by elementary calculus.

A Gaussian distribution $\mathcal{N}(\mu, \sigma^2)$ is parametrized by two parameters $\mu$ and $\sigma^2$. We write $x \sim \mathcal{N}(\mu, \sigma^2)$ to denote a real-valued random variable $x \in \mathbb{R}$ sampled by the (yet unspecified) Gaussian distribution. The paramters $\mu$ and $\sigma^2$ have simple interpretations.

**Fact 1.** *Let $x \sim \mathcal{N}(\mu, \sigma^2)$. Then the mean and variance of $x$ are*

$$\mathrm{E}[x] = \mu \text{ and } \mathrm{Var}[x] = \sigma^2.$$

We abbreviate $\mathcal{N} \stackrel{\text{def}}{=} \mathcal{N}(0, 1)$ for the special case of a Gaussian random variable with mean 0 and variance 1.

Some simple operations on Gaussian's produce new Gaussian's with their parameters naturally modified. First, scaling or shifting a Gaussian produces another Gaussian.

**Fact 2.** *Let $x \sim \mathcal{N}(0, \sigma^2)$ and let $\alpha \in \mathbb{R}$. Then*

$$\alpha x \sim \mathcal{N}(0, \alpha^2\sigma^2)$$

*and*

$$x + \alpha \sim \mathcal{N}(\alpha, \sigma^2).$$

Second, adding two Gaussians produces another Gaussian with the means and variances (necessarily) added together.

**Fact 3.** *Let $x_1 \sim \mathcal{N}(\mu_1, \sigma_1^2)$ and $x_2 \sim \mathcal{N}(\mu_2, \sigma_2^2)$, then $x_1 + x_2 \sim \mathcal{N}(\mu_1 + \mu_2, \sigma_1^2 + \sigma_2^2)$.*

We also note that Gaussian's have nice exponential moments. Recall that exponential moments previously appeared when developing the Chernoff bound. Likewise, the following fact will eventually imply (below) that sums of Gaussians squared are well concentrated.

**Fact 4.** *Let $x \in \mathcal{N}$ and $t < 1/2$. Then*

$$\mathrm{E}\left[e^{tx^2}\right] = \frac{1}{\sqrt{1-2t}}.$$

## 2.1 Concentration of length

We are also interested in ensembles of Gaussian random variables. For $k \in \mathbb{N}$, let $\mathcal{N}(\mu, \sigma^2)^k$ denote the distribution of $k$-dimensional vectors where each coordinate is a $k$-dimensional vector with unit length. That is, when we write $x \in \mathcal{N}^k(\mu, \sigma^2)$, we mean that each $x_i \sim \mathcal{N}(\mu, \sigma^2)$, independently. Note that for $x \in \mathcal{N}^k(0, \sigma^2)$ has expected squared length

$$\mathrm{E}\left[\|x\|^2\right] = \sum_{i=1}^{k} \mathrm{E}\left[x_i^2\right] = \sum_{i=1}^{k} \mathrm{Var}[x_i] = k\sigma^2.$$

As a direct consequence of Fact 4 above, the squared length $\|x\|^2$ of a Gaussian vector $x \sim \mathcal{N}^k$ will be extremely well concentrated, as follows.

**Fact 5.** *Let $x \sim \mathcal{N}(0, \sigma^2)^k$ be a Gaussian vector. Let $\alpha \geq 0$.*

1. *If $\alpha \leq 1$, then*

$$\mathrm{P}\left[\|x\|^2 \leq \alpha\, \mathrm{E}\left[\|x\|^2\right]\right] \leq \left(\alpha e^{1-\alpha}\right)^{k/2}.$$

*2. If $\alpha \geq 1$, then*

$$\mathrm{P}\Big[\|x\|^2 \geq \alpha\,\mathrm{E}\Big[\|x\|^2\Big]\Big] \leq \big(\alpha e^{1-\alpha}\big)^{k/2}.$$

*Proof.* Let us prove this fact because we only need the above facts to do so. Scaling $x$ (and invoking Fact 2), we can assume that $x \in \mathcal{N}^k$ and $\mu = \mathrm{E}\Big[\|x\|^2\Big] = k$.

Let $\alpha \in [0,1]$. For $t > 0$, we have

$$\mathrm{P}\Big[\|x\|^2 \leq \alpha k\Big] = \mathrm{P}\Big[e^{-t\|x\|^2} \geq e^{-t\alpha k}\Big] \overset{\text{(a)}}{\leq} \mathrm{E}\Big[e^{-t\|x\|^2}\Big] e^{t\alpha k} \overset{\text{(b)}}{=} \left(\prod_{i=1}^{k} \mathrm{E}\Big[e^{-tx_i^2}\Big]\right) e^{t\alpha k}$$
$$\overset{\text{(c)}}{=} \left(\frac{1}{1+2t}\right)^{k/2} \exp(\alpha t k) = \exp\left(k\left(\alpha t - \frac{1}{2}\ln(1+2t)\right)\right)$$

(a) is by Markov's inequality. (b) is by independence of the $x_i$'s (noting that $\|x\|^2 = \sum_i x_i^2$). (c) is by Fact 4. The (exponent of the) RHS is minimized by

$$\alpha = \frac{1}{1+2t} \iff t = \frac{1-\alpha}{2\alpha}. \tag{1}$$

Plugging in $t$ per (1) gives

$$\mathrm{P}\Big[\|x\|^2 \leq \alpha k\Big] \leq \big(\alpha e^{1-\alpha}\big)^{k/2},$$

as desired.

Now let $\alpha \geq 1$. For any $t \in (0, 1/2)$, we have

$$\mathrm{P}\Big[\|x\|^2 \geq \alpha k\Big] = \mathrm{P}\Big[e^{t\|x\|^2} \geq e^{t\alpha k}\Big] \overset{\text{(d)}}{=} \mathrm{E}\Big[e^{t\|x\|^2} e^{-t\alpha k}\Big]$$
$$\overset{\text{(e)}}{\leq} \left(\frac{1}{1-2t}\right)^{k/2} e^{-\alpha t k} = \exp\left(-\frac{k}{2}(2\alpha t + \ln(1-2t))\right)$$

by (d) Markov's inequality and (e) Fact 4. The RHS is minimized at

$$\alpha = \frac{1}{1-2t} \iff t = \frac{\alpha - 1}{2\alpha};$$

moreover, the RHS is in $(0, 1/2)$ for all $\alpha > 1$. Plugging in, we have

$$\mathrm{P}\Big[\|x\|^2 \geq \alpha k\Big] \leq \big(\alpha e^{1-\alpha}\big)^{k/2},$$

as desired. ∎

An important case is where $\alpha = (1 \pm \epsilon)$ and $\epsilon > 0$ is close to 0. Then Fact 5 implies the following.

**Lemma 2.** *Let $x \sim \mathcal{N}\big(0, \sigma^2\big)^k$ be a Gaussian vector. Let $\epsilon > 0$ be sufficiently small. Then*

$$\mathrm{P}\Big[\|x\|^2 \leq (1-\epsilon)\,\mathrm{E}\Big[\|x\|^2\Big]\Big] \leq e^{-\epsilon^2 k/8}.$$

*and*

$$\mathrm{P}\Big[\|x\|^2 \geq (1+\epsilon)\,\mathrm{E}\Big[\|x\|^2\Big]\Big] \leq e^{-\epsilon^2 k/8}$$

*Proof.* By scaling, we can assume that $\sigma^2 = 1$ and $\|x\|^2 = 1$. We have

$$\mathrm{P}\Big[\|x\|^2 \geq (1+\epsilon)k\Big] \overset{\text{(a)}}{\leq} \big((1+\epsilon)e^{-\epsilon}\big)^{k/2} \overset{\text{(b)}}{\leq} e^{.15\epsilon^2 k}.$$

Here (a) is by Fact 5. (b) is because

$$1 + \epsilon \leq e^{\epsilon - \frac{1}{4}\epsilon^2}.$$

for $\epsilon$ sufficiently small. Likewise, we have

$$\mathrm{P}\Big[\|x\|^2 \leq (1-\epsilon)k\Big] \overset{\text{(c)}}{\leq} ((1-\epsilon)e^{\epsilon})^{k/2} \overset{\text{(d)}}{\leq} e^{-\epsilon^2 k/8}.$$

(c) is by Fact 5. (d) is because

$$1 - \epsilon \leq e^{-\epsilon - \frac{1}{4}\epsilon^2}$$

for $\epsilon > 0$ sufficiently small. ■

# 3 Random Projections

So far, we know that Gaussian random variables can be scaled and added together, and that the length of a squared Gaussian vector is well concentrated around its expectation. In fact this is all we need for the dimensionality reduction result mentioned above. The first lemma considers the projection of a single vector.

**Lemma 3.** *Let $A \in \mathcal{N}^{k \times d}$ be a random matrix where each coordinate $A_{ij}$ is an independently drawn sample from $\mathcal{N}$. Let $\epsilon > 0$ be sufficiently small. For any vector $x$,*

$$\mathrm{P}\left[(1-\epsilon)\|x\|^2 \leq \frac{1}{k}\|Ax\|^2 \leq (1+\epsilon)\|x\|^2\right] \geq 1 - 2e^{-k/8}.$$

*Proof.* Scaling if necessary, we can assume without loss of generality that $\|x\| = 1$. For $i \in [k]$, let $a_i = A^T e_i$ be the $i$th row of $A$. We have $a_i \sim \mathcal{N}^n$. Consider $\langle a_i, x\rangle = (Ax)_i$, as a random variable. By facts 2 and 3, $\langle a_i, x\rangle$ is a Gaussian random variable with mean 0 and variance

$$\sum_{j=1}^{n} \mathrm{Var}[x_j A_{ij}] = \sum_{j=1}^{n} x_j^2 = 1.$$

That is, $\langle a_i, x\rangle \sim \mathcal{N}$ for each $i$. In turn, we have $(Ax) \sim \mathcal{N}^k$. As a $k$-dimensional Gaussian vector, $\|Ax\|^2$ will be very well concentrated at its mean per Fact **??**. ■

Consider Theorem 1 from the introduction, where we have a set of $n$ points $P \subseteq \mathbb{R}^d$, and randomly project it into $\mathbb{R}^k$ with the linear function

$$f(x) = \frac{1}{\sqrt{k}} Ax,$$

where $k = O\big(\log(n)/\epsilon^2\big)$ and $A \sim \mathcal{N}^{k \times d}$. By Lemma 3, for each $x \in P$, we have

$$(1-\epsilon)\|x\|^2 \leq \|f(x)\|^2 \leq (1+\epsilon)\|x\|^2 \tag{2}$$

with probability of error (say) $\leq 1/n^{10}$. By the union bound, we have (2) for all $x$ with probability of error $\leq 1/n^9$. Theorem 1 also promised that all pairwise distances are preserved up to an $(1 \pm \epsilon)$-multiplicative factor. Here we take advantage of the *linearity* of $f$. We have

$$\|f(x) - f(y)\|^2 = \|f(x - y)\|^2$$

for any two points $x, y \in P$. We now argue, as before, that the length of pairwise differences $x - y$ is preserved with high probability.

# 4 Gaussians

Based on Theorem 1, a distribution satisfying facts 1, 2, 3, and 4 (from which all other facts and theorems are derived) may seem too good to be true. Let us now define this distribution formally and verify these simple facts.

The **Gaussian or normal distribution with mean $\mu \in \mathbb{R}$ and variance $\sigma^2 \geq 0$** is the real-valued random variable with density function

$$f(x) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2}. \tag{3}$$

(By Lemma 4 below, this random variable indeed has mean $\mu$ and variance $\sigma^2$.) We let $\mathcal{N}(\mu, \sigma^2)$ to denote the normal distribution with mean $\mu$ and variance $\sigma^2$, and write $X \sim \mathcal{N}(\mu, \sigma^2)$ to denote a random variable $X \in \mathbb{R}$ with distribution $\mathcal{N}(\mu, \sigma^2)$. A **normalized Gaussian** or **standard normal** random variable is a Gaussian random variable with mean 0 and variance 1. We abbreviate $\mathcal{N}(0, 1)$ by $\mathcal{N}$. For $n \in \mathbb{N}$, we let $\mathcal{N}^n$ denote the joint distribution of $n$ independent normalized Gaussian random variables.

## 4.1 Some preliminary calculus

**Lemma 4.** *Let $\mu \in \mathbb{R}$, $\sigma > 0$, and*

$$f(x) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-(x-\mu)^2/2\sigma^2}.$$

*Then we have the following.*

1. $\displaystyle\int_{-\infty}^{\infty} f(x) = 1.$
2. $\displaystyle\int_{-\infty}^{\infty} x f(x) = \mu.$
3. $\displaystyle\int_{-\infty}^{\infty} (x - \mu)^2 f(x) = \sigma^2.$

*Proof.* We consider the normalized case $\mu = 0$ and $\sigma = 1$. The general case follows by appropriate change of variables. We have

$$\begin{aligned}\left(\int_{-\infty}^{\infty} e^{-x^2/2}\, dx\right)^2 &= \int_{-\infty}^{\infty} e^{-(x^2+y^2)/2}\, dx dy = \int_0^{2\pi} \int_0^{\infty} r e^{-r^2/2}\, dr d\theta \\ &= 2\pi \int_0^{\infty} r e^{-r^2/2}\, dr = 2\pi.\end{aligned}$$

Taking the square root of both sides gives the first claim. For the second claim, we have

$$\int_{-\infty}^{\infty} xe^{-x^2/2}\,dx = \left[e^{-x^2/2}\right]_{-\infty}^{+\infty} = 0.$$

For the third claim, we have

$$\int_{-\infty}^{\infty} x^2 e^{-x^2/2}\,dx \overset{\text{(a)}}{=} \left[-xe^{-x^2/2}\right]_{-\infty}^{\infty} + \int_{-\infty}^{\infty} e^{-x^2/2} = 1$$

by (a) integration by parts. ■

Lemma 4 immediately implies both Fact 1 and Fact 2, which we restate for convenience.

**Fact 1.** *Let $x \sim \mathcal{N}(\mu, \sigma^2)$. Then the mean and variance of $x$ are*

$$\mathrm{E}[x] = \mu \text{ and } \mathrm{Var}[x] = \sigma^2.$$

**Fact 2.** *Let $x \sim \mathcal{N}(0, \sigma^2)$ and let $\alpha \in \mathbb{R}$. Then*

$$\alpha x \sim \mathcal{N}(0, \alpha^2\sigma^2)$$

*and*

$$x + \alpha \sim \mathcal{N}(\alpha, \sigma^2).$$

## 5 Rotational symmetry of Gaussian vectors

Let $X \sim \mathcal{N}^n$, and let $f(x)$ be the density function of $x$. The density function has the following compact form. The key feature is that the density at a point it only depends on the squared length of the point. That is, it is *rotationally symemetric*.

**Lemma 5.** $f(x) = (2\pi)^{-n/2} e^{-\langle x,x\rangle/2}$.

*Proof.* Since each $x_i \sim \mathcal{N}$ independently, we have

$$f(x) = \prod_{i=1}^{n} (2\pi)^{-1/2} e^{-x_i^2/2} = (2\pi)^{-n/2} e^{-\langle x,x\rangle/2},$$

as desired. ■

**Lemma 6.** *For any orthonormal matrix $U$, and random Gaussian vector $x$, $Ux \sim \mathcal{N}^n$.*

*Proof.* $U$ induces a rotation, and the Gaussian is rotationally symmetric. For those who prefer explicit calculations, we have

$$f(Ux) \overset{\text{(a)}}{=} (2\pi)^{-n/2} e^{-\langle Ux,Ux\rangle/2} \overset{\text{(b)}}{=} (2\pi)^{-n/2} e^{-\langle x,x\rangle/2} \overset{\text{(c)}}{=} f(x),$$

where (a) is by Lemma 5, (b) is because $U^T U = I$, and (c) is by Lemma 5. ■

**Lemma 7.** *Let $x \sim \mathcal{N}^n$ and $u \in \mathbb{R}^n$. Then $\langle u, x\rangle \sim \mathcal{N}\left(0, \|u\|^2\right)$.*

*Proof.* It suffices to assume $u$ is a unit vector. By extending $u$ to an orthonormal basis, let $u = U^T e_1$ for an orthogonal matrix $U$. Then

$$\langle u, x\rangle = \left\langle U^t e_1, x\right\rangle = \langle e_1, Ux\rangle = (Ux)_1 \overset{\text{(a)}}{\sim} \mathcal{N},$$

where (a) is by Lemma 6. ■

Fact 3, which says that Gaussians sum nicely, now follows by combination of Fact 2 and Lemma 7. We restate Fact 3 for convenience and leave the proof to the reader.

**Fact 3.** *Let $x_1 \sim \mathcal{N}(\mu_1, \sigma_1^2)$ and $x_2 \sim \mathcal{N}(\mu_2, \sigma_2^2)$, then $x_1 + x_2 \sim \mathcal{N}(\mu_1 + \mu_2, \sigma_1^2 + \sigma_2^2)$.*

## 6 Moments of squared Gaussian random variables

The last fact to probe concerns the moment generating function of the square of a Gaussian random variable. Recall that amplifying the following bound leads to the concentration of length of high-dimensional Gaussian vectors, which in turn, allows us to obliviously embed high-dimensional data in Theorem 1.

**Fact 4.** *Let $x \in \mathcal{N}$ and $t < 1/2$. Then*

$$\mathrm{E}\Big[e^{tx^2}\Big] = \frac{1}{\sqrt{1-2t}}.$$

*Proof.* We have

$$\begin{aligned}\mathrm{E}\Big[e^{tx^2}\Big] &= \int_s e^{ts^2}\mathrm{P}[x=s] \overset{\text{(a)}}{=} \frac{1}{\sqrt{2\pi}}\int_s e^{(2t-1)s^2/2} \\ &= \frac{1}{\sqrt{2\pi}}\int_s e^{-s^2/2\sigma^2} \text{ for } \sigma = 1/\sqrt{1-2t} \\ &\overset{\text{(b)}}{=} \frac{1}{\sqrt{1-2t}}.\end{aligned}$$

Here (a) plugs in the density function from equation (3). (b) is by Lemma 4.1 w/r/t the density function for $\mathcal{N}\big(0,\sigma^2\big)$. ■

## 7 Exercises

**Exercise 1.** Using only Fact 2, show that for $x \sim \mathcal{N}\big(\mu,\sigma^2\big)$ and $\alpha \in \mathbb{R}$,

$$\alpha x \sim \mathcal{N}\big(\alpha\mu, \alpha^2\sigma^2\big).$$

**Exercise 2.** Show that there exists universal constants $c_1, c_2 > 0$ such that for all $x$ with $|x| \le c_1$,

$$1 + x \le e^{x - c_2 x^2}.$$

(In other words, you can choose whatever constants $c_1$ and $c_2$ are convenient to you.)

**Exercise 3.** Let $P \subseteq \mathbb{R}^d$ be a set of $n$ points. Let $f : \mathbb{R}^d \to \mathbb{R}^k$ be a random projection with $k = O(\log n)$ (per Theorem 1). Recall that with high probability (say, $\ge 1 - 1/n^4$), we have

$$(1-\epsilon)\|x\|^2 \le \|f(x)\|^2 \le (1+\epsilon)\|x\|^2$$

for all $x \in P$, and we also have

$$(1-\epsilon)\|x-y\|^2 \le \|f(x)-f(y)\|^2 \le (1+\epsilon)\|x-y\|^2$$

as well as

$$(1-\epsilon)\|x+y\|^2 \le \|f(x)+f(y)\|^2 \le (1+\epsilon)\|x+y\|^2$$

for all $x, y \in P$. Show that with high probability, we also have

$$|\langle f(x), f(y)\rangle - \langle x, y\rangle| \le \frac{\epsilon}{2}\Big(\|x\|^2 + \|y\|^2\Big)$$

for all $x, y \in P$.

*Hint: What is $\|x+y\|^2 - \|x-y\|^2$?*

## References


[1] William Johnson and Joram Lindenstrauss. "Extensions of Lipschitz maps into a Hilbert space". In: *Contemporary Mathematics* 26 (Jan. 1984), pp. 189–206.

[2] Dimitris Achlioptas. "Database-friendly random projections". In: *Proceedings of the Twentieth ACM SIGACT-SIGMOD-SIGART Symposium on Principles of Database Systems, May 21-23, 2001, Santa Barbara, California, USA*. Ed. by Peter Buneman. ACM, 2001.

[3] Nir Ailon and Bernard Chazelle. "The Fast Johnson–Lindenstrauss Transform and Approximate Nearest Neighbors". In: *SIAM J. Comput.* 39.1 (2009), pp. 302–322. DOI: 10.1137/060673096. URL: https://doi.org/10.1137/060673096.

[4] Anirban Dasgupta, Ravi Kumar, and Tamás Sarlós. "A sparse Johnson: Lindenstrauss transform". In: *Proceedings of the 42nd ACM Symposium on Theory of Computing, STOC 2010, Cambridge, Massachusetts, USA, 5-8 June 2010*. Ed. by Leonard J. Schulman. ACM, 2010, pp. 341–350.

[5] Tomas Mikolov, Ilya Sutskever, Kai Chen, Gregory S. Corrado, and Jeffrey Dean. "Distributed Representations of Words and Phrases and their Compositionality". In: *Advances in Neural Information Processing Systems 26: 27th Annual Conference on Neural Information Processing Systems 2013. Proceedings of a meeting held December 5-8, 2013, Lake Tahoe, Nevada, United States*. Ed. by Christopher J. C. Burges, Léon Bottou, Zoubin Ghahramani, and Kilian Q. Weinberger. 2013, pp. 3111–3119. URL: http://papers.nips.cc/paper/5021-distributed-representations-of-words-and-phrases-and-their-compositionality.

[6] Daniel M. Kane and Jelani Nelson. "Sparser Johnson-Lindenstrauss Transforms". In: *J. ACM* 61.1 (2014), 4:1–4:23.

[7] David P. Woodruff. "Sketching as a Tool for Numerical Linear Algebra". In: *Found. Trends Theor. Comput. Sci.* 10.1-2 (2014), pp. 1–157.

[8] Petros Drineas and Michael W. Mahoney. "Lectures on Randomized Numerical Linear Algebra". In: *CoRR* abs/1712.08880 (2017). arXiv: 1712.08880. URL: http://arxiv.org/abs/1712.08880.

[9] Michael B. Cohen, T. S. Jayram, and Jelani Nelson. "Simple Analyses of the Sparse Johnson-Lindenstrauss Transform". In: *1st Symposium on Simplicity in Algorithms, SOSA 2018, January 7-10, 2018, New Orleans, LA, USA*. Ed. by Raimund Seidel. Vol. 61. OASICS. Schloss Dagstuhl - Leibniz-Zentrum für Informatik, 2018, 15:1–15:9. URL: https://doi.org/10.4230/OASIcs.SOSA.2018.15.

[10] Roman Vershynin. *High-Dimensional Probability: An Introduction with Applications in Data Science*. Cambridge Series in Statistical and Probabilistic Mathematics. Cambridge University Press, 2018. DOI: 10.1017/9781108231596.