Princeton Univ. F'20 COS 521: Advanced Algorithm Design

Lecture 12: Dimensionality Reduction and the Johnson-Lindenstrauss Lemma

Lecturer: *Sahil Singla* Last updated: October 5, 2020

# 1 Preliminaries

Very high-dimensional vectors are ubiquitous in science, engineering, and machine learning. They give a simple way of representing data: for each object we want to study, we collect a very large set of numerical parameters, often with no inherent order or structure. We use these parameters to compare, analyze, and make inferences about those objects.

High-dimensional data comes from genetic data sets, time series (e.g. audio or seismographic data), image data, etc. It is also a common output of feature generation algorithms.

Feature generation algorithms are commonly used to pre-process image and audio data as well. For example, Shazam and other "song matching" services preprocess audio by computing a spectrogram, which essentially computes many Fourier transforms of different sections of the signal, shifted to start at different time points. More on this example later.

What do we want to do with such high dimensional vectors? Cluster them, use them in regression analysis, feed them into machine learning algorithms. As an even more basic goal, all of these tasks require being able to determine if one vector is similar to another. Even this simple task becomes an unwiedly in high-dimensions.

# 2 Dimensionality Reduction

The goal of dimensionality reduction is to reduce the cost of working with high-dimensional data by representing it more compactly. Instead of working with an entire vector, can we find a more compact "fingerprint" – i.e. a shorter vector – that at least allows us to quickly compare vectors? Or maybe the fingerprint preserves certain properties of the original vector that allows it to be used in other downstream tasks.

Computer scientists have developed a remarkably general purpose toolkit of dimensionality reduction methods for constructing compact representations that can be used effectively in a huge variety of downstream tasks. In this section of the course, we will study some of those methods.

# 3 The Johnson-Lindenstrauss Lemma

We start with a particular powerful and influential result in high-dimensional geometry. It applies to problems involving the $\ell_2$ norm:

$$\|x\|_2 = \sqrt{\sum_{i=1}^{m} x_i^2}$$

For two vectors $x$ and $y$, $\|x-y\|_2$ is the Euclidean distance.

**Problem 1.** *Given $n$ points $v^1, v^2, ..., v^n \in \mathbb{R}^d$, we want to find a function $f : \mathbb{R}^d \to \mathbb{R}^m$ such that $m$ is much smaller than $d$ and for all $i, j$,*

$$(1-\epsilon)\|v^i - v^j\|_2 \leq \|f(v^i) - f(v^j)\|_2 \leq (1+\epsilon)\|v^j - v^j\|_2. \tag{1}$$

*In other words, the distance between all pairs of points in preserved.*

The following main result (Lemma in their words) is by Johnson & Lindenstrauss [1]:

**Theorem 2** (Johnson-Lindenstrauss Lemma)**.** *There is a function $f$ satisfying (1) that maps vectors to $m = O(\frac{\log n}{\epsilon^2})$ dimensions. In fact, $f$ is a* linear mapping *and can be applied in a computationally efficient way!*

The following ideas do not work to prove this theorem: (a) take a random sample of $m$ coordinates out of $d$. (b) Partition the $d$ coordinates into $m$ subsets of size about $n/m$ and *add* up the values in each subset to get a new coordinate.[1]

We're going to choose $f$ randomly. In particular, let $G$ be a $m \times d$ *random matrix* with each entry a normal random variable, $G_{i,j} \sim \mathcal{N}(0,1)$. Let $\Pi = \frac{1}{\sqrt{m}}G$:

$$f(x) = \Pi x.$$

So each entry in $u = f(v)$ equals $v \cdot g$ for some vector $g$ filled with scaled Gaussian random variables. Other choices for $G$ work: for example, we can use random signs or a random orthonormal matrix (used in the original proof). More on this next lecture.

We're going to prove a slightly stronger statement for this map:

**Theorem 3** ($(\epsilon, \delta)$-JL property)**.** *If $m = O(\log(1/\delta)/\epsilon^2)$, then for any vector $x$,*

$$(1-\epsilon)\|x\|_2^2 \leq \|\Pi x\|_2^2 \leq (1+\epsilon)\|x\|_2^2 \tag{2}$$

*with probability $(1-\delta)$.*

Note that, while stated with the squared Euclidean norm, (2) immediately implies that $(1-\epsilon)\|x\|_2 \leq \|\Pi x\|_2 \leq (1+\epsilon)\|x\|_2$ (just by taking a square root of all sides, and observing that this brings the constants closer to 1). Then, to prove Theorem 2 from this stronger statement, we use the linearity of $f$ to see that:

$$\|f(v^i) - f(v^j)\|_2 = \|\Pi v^i - \Pi v^j\|_2 = \|\Pi(v^i - v^j)\|_2.$$

So, with probability $(1-\delta)$ we preserve one distance. We have $\binom{n}{2} = O(n^2)$ distances total. By a union bound, we preserve all of them with probability $1-\delta$ as long as we reduce $\delta$ to $\delta/\binom{n}{2}$, which means that $m = O(\log(n/\delta)/\epsilon^2)$. This gives Theorem 2. So, we can focus our attention on proving Theorem 3.

[1] To see why these approaches fail whp, consider the case of two vectors: $(1, 0, \ldots, 0)$ and $(0, 1, 0, \ldots, 0)$. Then the first approach succeeds iff we happen to pick coordinate one or two as one of the coordinates, which is unlikely. To see why the second approach fails, consider two vectors $(1, \ldots, 1, 0, \ldots, 0)$ and $(0, \ldots, 0, 1, \ldots, 1)$. Then the second approach whp generates nearly-identical vectors even though the initial two vectors are far apart.

*Proof.* Let $w = Gx$ be a scaling of our dimension reduced vector. Our goal is to show that $\|x\|_2^2$ is approximated by:

$$\|\Pi x\|_2^2 = \|\frac{1}{\sqrt{m}} Gx\|_2^2 = \frac{1}{m}\sum_{i=1}^{m} w_i^2.$$

Consider one term of the sum, $w_i^2$, which is a random variable since $G$ is chosen randomly. We will start by showing that each term is equal to $\|x\|_2^2$ in expectation. We have:

$$w_i = \sum_{j=1}^{d} x_j g_j$$

where each $g_j \sim \mathcal{N}(0,1)$. So $\mathbb{E}[w_i] = \sum_{j=1}^{d} x_j \mathbb{E}[g_j] = 0$ and thus $\mathrm{Var}[w_i] = \mathbb{E}[w_i^2]$. It follows that:

$$\mathbb{E}[w_i^2] = \mathrm{Var}[w_i] = \sum_{j=1}^{d} \mathrm{Var}[x_j g_j] = \sum_{j=1}^{d} x_j^2 \,\mathrm{Var}[g_j] = \sum_{j=1}^{d} x_j^2 = \|x\|_2^2.$$

Thus $\mathbb{E}[w_i^2] = \|x\|_2^2$ and our estimate is correct in expectation:

$$\mathbb{E}\left[\frac{1}{m}\sum_{i=1}^{m} w_i^2\right] = \|x\|_2^2.$$

How do we know that it's close to this expectation with high probability? We actually know that $w_i$ is a *normal random variable.*

**Fact 4** (Stability of Gaussian random variables)**.** *If $X$ and $Y$ are independent and $X \sim \mathcal{N}(0, a^2)$ and $Y \sim \mathcal{N}(0, b^2)$, then $X + Y \sim \mathcal{N}(0, a^2 + b^2)$. The property that the sum of Gaussian's remains Gaussian is known as "stability"*[2]*.*

So $w_i \sim \mathcal{N}(0, \|x\|^2) = \|x\|_2 \cdot \mathcal{N}(0,1)$. It follows that $w_i^2$ is a $\chi^2$ (chi-squared) random variable and $\frac{1}{m}\sum_{i=1}^{m} w_i^2$ is a chi-squared random variable with $m$ degrees of freedom. You can look up the CDF on Wikipedia for a $\chi^2$ tail bound, but it essentially concentrates around its mean as well as a Gaussian. In particular, if $v = \frac{1}{m}\sum_{i=1}^{m} w_i^2$, then[3]:

$$\Pr\left[|\mathbb{E}v - v| \geq \epsilon \mathbb{E}v\right] \leq 2e^{-m\epsilon^2/8}.$$

So, if we set $m = O(\log(1/\delta)/\epsilon^2)$ then $\|\Pi x\|_2^2 = \frac{1}{m}\sum_{i=1}^{m}$ satisfies:

$$\|x\|_2^2 - \epsilon\|x\|_2^2 \leq \|\Pi x\|_2^2 \leq \|x\|_2^2 + \epsilon\|x\|_2^2$$

with probability $1 - \delta$. □

[2]There are other classes of stable distributions, but the normal distribution is the only stable distribution with bounded variance, which gives some intuition for why the central limit theorem holds for random variables with bounded variance.

[3]See e.g. https://www.stat.berkeley.edu/~mjwain/stat210b/Chap2_TailBounds_Jan22_2015.pdf

**Theorem 5** (Alternative $(\epsilon, \delta)$-JL construction)**.** *If $m = O(\log(1/\delta)/\epsilon^2)$, then for any vector $x$,*

$$(1-\epsilon)\|x\|_2^2 \leq \|\Pi x\|_2^2 \leq (1+\epsilon)\|x\|_2^2 \tag{3}$$

*with probability $(1-\delta)$.*

*Proof.* Choose $m$ vectors $z^1, ..., z^m \in \Re^d$ at random by choosing each coordinate randomly from $\{\sqrt{\frac{1}{m}}, -\sqrt{\frac{1}{m}}\}$. Then consider the mapping from $\Re^d$ to $\Re^m$ given by

$$x \longrightarrow (x \cdot z^1, x \cdot z^2, \ldots, x \cdot z^m).$$

Observe that this is the same as letting $\Pi$ be a random matrix where each entry is independently uniform in $\{\sqrt{1/m}, -\sqrt{1/m}\}$. Now, consider a vector $x \in \mathbb{R}^d$, and let's consider the vector $u = \Pi x$:

$$\|u\|^2 = \sum_{k=1}^{m} \left(z^k \cdot x\right)^2 = \sum_{k=1}^{m} \left(\sum_{\ell=1}^{d} z_\ell^k x_\ell\right)^2.$$

Again, we want to compute the expected value of one term $\left(\sum_{\ell=1}^{d} z_\ell^k x_\ell\right)^2$ and argue that it is exactly $\|x\|_2^2/m$. Indeed:

$$\mathbb{E}[(\sum_{\ell=1}^{d} z_\ell^k x_\ell)^2] = \mathbb{E}[\sum_{\ell=1}^{d}\sum_{j=1}^{d} z_\ell^k x_\ell z_j^k x_j]$$

$$= \sum_{\ell=1}^{d}\sum_{j=1}^{d} x_\ell x_j \mathbb{E}[z_\ell^k z_j^k] = \sum_{\ell=1}^{d} x_\ell^2 \mathbb{E}[(z_\ell^k)^2]$$

$$= ||x||_2^2 \cdot \frac{1}{m}.$$

Above, the first two equalities are simply expanding linearity of expectation. The penultimate equality observes that $\mathbb{E}[x_\ell^k x_j^k] = 0$ whenever $\ell \neq k$, and the final equality observes that $\mathbb{E}[(x_\ell^k)^2] = \frac{1}{m}$.

Therefore, the expectation of $\|u\|^2$ is $\|x\|^2$. If we show that $\|u\|^2$ is concentrated enough around its mean, then it would prove the theorem. More formally, this is done in the following Chernoff bound lemma. The point is that because each $u_\ell^2$ is independent and bounded, their sum should concentrate around its expectation.

Seemingly, the the right proof approach should be to bound the random variables and then use a Chernoff bound, or perhaps Bernstein's inequality. Unfortunately, this doesn't give a particularly good bound because the random variables can be quite large, albeit with really tiny probability (in particular, observe that each $u_k^2$ could be as large as $|x|_1^2/m$). The proof of the following lemma is omitted:

**Lemma 6.** *There exists a constant $c$ such that:*

$$\Pr[||u||_2^2 \notin (||x||_2^2 \pm \varepsilon||x||_2^2)] \leq e^{-c\varepsilon^2 m}.$$

With Lemma 6, we can now observe that taking $m = \ln(1/\delta)/(c\varepsilon^2) = O(\log(1/\delta)/\varepsilon^2)$ results in a failure probability of only $\delta$. □

It's worth noting that Theorem 2 is tight – i.e. there are point sets that cannot be embedded into less than $O(\log n/\epsilon^2)$ dimensions if we want to preserve all pairwise distances. This was proven up to a $\log(1/\epsilon)$ factor by Noga Alon in [2]. The fully tight result was only obtained in 2017 [3]. The result was proven first for *linear embeddings* and then extended to a lower-bound for all possible functions $f$.

# 4 Applications

There are many, many applications of the JL lemma. Here are a few that we will see on the problem set or in later classes:

- Approximate all-pairs distances in $O(n^2 \log n + nd)$ time vs. the naive $O(n^2 d)$ time.
- Approximate distance based clustering.
- Approximate support vector machine (SVM) classification and more.
- Sparse recovery.compressed sensing.
- Approximate linear regression.

## 4.1 Linear regression

In addition to its use in proving the original lemma about distances, the $(\epsilon, \delta)$-JL property for norm preservation is often directly useful in applications. Furthermore, many applications crucially use the *linearity* of the Johnson-Lindenstrauss embedding, not just its approximation properties. Here we consider a classic example: least squares regression.

Given $n$ data vectors $a_1, \ldots, a_n \in \mathbb{R}^d$ and $n$ response values $y_1, \ldots, y_n \in \mathbb{R}$. Usually we think of $a_1, \ldots, a_n$ as the rows in an $n \times d$ matrix $A$ and $y_1, \ldots, y_n$ as entries in $n$ length vector $y$. Goal:

$$\min_{x \in \mathbb{R}^d} \sum_{i=1}^{n} (a_i \cdot x - y_i)^2 = \min_{x \in \mathbb{R}^d} \|Ax - y\|_2^2 \tag{4}$$

Typically this probably requires $O(nd^2)$ time to solve. We will speed this up by reducing $n$ using the Johnson-Lindenstrauss Lemma.[4] In particular, let $\Pi \in \mathbb{R}^{m \times n}$ be chosen from a random family of matrices satisfying Theorem 3. To obtain an approximate solution we will solve the "sketched" problem:

$$\min_{x \in \mathbb{R}^d} \|\Pi A x - \Pi y\|_2^2, \tag{5}$$

[4]Importantly, note that we are aiming to reduce $n$ (the number of vectors) and not $d$ (the dimension). Of course, since we only care about the matrix $A$, we could think of $n$ as the dimension and $d$ as the number of vectors, but just be aware that this deviates from their semantic meaning.

which can be solved in $O(md^2)$ time (once $\pi A$ and $\Pi y$ are computed — we won't discuss this aspect, but there are JL transforms which are also fast). We want to prove that a solution to this smaller problem is a good approximate solution to the original. Before doing so, we claim a simpler result:

**Lemma 7.** *As long as $m = O(\log(1/\delta)/\epsilon^2)$ then, for any particular $x$,*

$$(1-\epsilon)\|Ax-y\|_2^2 \leq \|\Pi Ax - \Pi y\|_2^2 \leq (1+\epsilon)\|Ax-y\|_2^2$$

*with probability $1-\delta$.*

This is a direct consequence of Theorem 3, applied to the vector $Ax - y$.

If we could show the same result *for all* $x$ then we would be in good shape. Specifically, let $x^*$ be the optimal solution for the original regression problem (4) and let $\tilde{x}^*$ be the optimal solution for the sketched problem (5). We have:

$$\|A\tilde{x}^* - y\|_2^2 \leq \frac{1}{1-\epsilon}\|\Pi A\tilde{x}^* - \Pi y\|_2^2 \leq \frac{1}{1-\epsilon}\|\Pi Ax^* - \Pi y\|_2^2 \leq \frac{1+\epsilon}{1-\epsilon}\|Ax^* - y\|_2^2$$

For $\epsilon \leq .25$, $\frac{1+\epsilon}{1-\epsilon} \leq 1+3\epsilon$. So we would get a relative error approximation to the regression problem. Question: For the argument above, why did we need a bound *for all* $x$?[5]

But how would we extend Lemma 7 to all $x$? We certainly can't use a union bound argument – there are an infinite number of possible vectors $x$.

## 5 Beyond the Union Bound

Recall that we have some $A \in \mathbb{R}^{n\times d}$ and some $y \in \mathbb{R}^n$ (we have $n$ training examples for our linear regression, each in $\mathbb{R}^d$), we want to approximately solve:

$$\min_{x\in\mathbb{R}^d} \|Ax-y\|_2^2 \tag{6}$$

by instead solving the "sketched" problem

$$\min_{x\in\mathbb{R}^d} \|\Pi Ax - \Pi y\|_2^2. \tag{7}$$

As long as $\Pi$ is chosen so that $m \leq n$, then $\Pi A$ contains fewer data points than $A$ and (7) can be solved much faster than (6): in $O(md^2)$ vs. $O(nd^2)$ time.

Let $\tilde{x}^*$ be the optimal solution for (7). We want to argue that

$$\|A\tilde{x}^* - y\|_2^2 \leq (1+\epsilon)\min_{x\in\mathbb{R}^d}\|Ax-y\|_2^2,$$

and saw that, to do so, it suffices to prove:

$$\forall x \in \mathbb{R}^d \qquad (1-\epsilon)\|Ax-y\|_2^2 \leq \|\Pi(Ax-y)\|_2^2 \leq (1+\epsilon)\|Ax-y\|_2^2. \tag{8}$$

Proving this statement requires establishing a Johnson-Lindenstrauss type bound for an *infinity* of possible vectors $Ax - y$, which obviously can't be tackled with a union bound argument. Today we will see how to prove this result using a different approach.

[5] Answer: Because the solution $\tilde{x}^*$ *depends on* $\Pi$. So we cannot simply fix $\tilde{x}^*$ and then use Theorem 3, because then we won't have the right $\tilde{x}^*$.

## 6 Subspace Embeddings

We will prove a more general statement that implies (8) and is useful in other applications.

**Theorem 8.** *Let $\mathcal{U} \subset \mathbb{R}^n$ be a $d$-dimensional linear subspace in $\mathbb{R}^n$. If $\Pi \in \mathbb{R}^{m \times n}$ is chosen from any distribution $\mathcal{D}$ satisfying Theorem 3, then with probability $1 - \delta$,*

$$(1-\epsilon)\|v\|_2 \leq \|\Pi v\|_2 \leq (1+\epsilon)\|v\|_2 \tag{9}$$

*for* all $v \in \mathcal{U}$, *as long as* $m = O\left(\frac{d\log(1/\epsilon)+\log(1/\delta)}{\epsilon^2}\right)$[6].

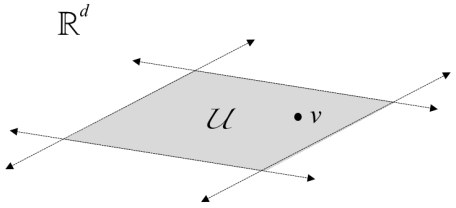


Figure 1: Theorem 8 extends Theorem 3 to all points in a linear subspace $\mathcal{U}$.

How does Theorem 8 imply (8)? We can apply it to the $d+1$ dimensional subspace spanned by $A$'s $d$ columns and $y$. Every vector $Ax - y$ lies in this subspace. So, for regression, we will require dimension $m = O\left(\frac{(d+1)\log(1/\epsilon)}{\epsilon^2}\right)$. In particular, note that this means we can approximately solve linear regression over $n \gg d$ examples for the same work as $m = O_\varepsilon(d)$ examples.

We start with the observation that Theorem 8 holds as long as (9) holds for all points on the unit sphere in $\mathcal{U}$. This is a consequence of linearity. We denote the sphere $S_\mathcal{U}$:[7]

$$S_\mathcal{U} = \{v \,|\, v \in \mathcal{U} \text{ and } \|v\|_2 = 1\}.$$

Any point $v \in \mathcal{U}$ can be written as $cx$ for some scalar $c$ and some point $x \in S_\mathcal{U}$. If $(1-\epsilon)\|x\|_2 \leq \|\Pi x\|_2 \leq (1+\epsilon)\|x\|_2$ then $c(1-\epsilon)\|x\|_2 \leq c\|\Pi x\|_2 \leq c(1+\epsilon)\|x\|_2$ and thus $(1-\epsilon)\|cx\|_2 \leq \|\Pi cx\|_2 \leq (1+\epsilon)\|cx\|_2$.

## 7 An argument via $\epsilon$-nets

We will prove Theorem 8 by showing that there exists a large but *finite* set of points $N_\epsilon \subset S_\mathcal{U}$ such that, if (9) holds for all $v \in N_\epsilon$, then it must hold for all $v \in S_\mathcal{U}$, and by the argument above, for all $v \in \mathcal{U}$. $N_\epsilon$ is called an "$\epsilon$-net".

**Lemma 9.** *For any $\epsilon \leq 1$, there exists a set $N_\epsilon \subset S_\mathcal{U}$ with $|N_\epsilon| = \left(\frac{4}{\epsilon}\right)^d$ such that $\forall v \in S_\mathcal{U}$,*

$$\min_{x \in N_\epsilon} \|v - x\| \leq \epsilon.$$

[6] It's possible to obtain a slightly tighter bound of $O\left(\frac{d+\log(1/\delta)}{\epsilon^2}\right)$. It's a nice challenge to try proving this. Hint: use a constant factor net $N_{O(1)}$ instead of an $\epsilon$ net $N_\epsilon$ as we do below.

[7] Below, write the vectors $v$ using any basis for $\mathcal{U}$, and let their norm be their norm written in this basis.

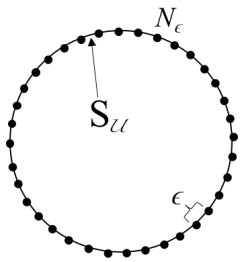


Figure 2: An $\epsilon$-net $N_\epsilon$ for a sphere in a 2 dimensional subspace $\mathcal{U}$.

### Construction of the $\epsilon$-net.

*Proof.* Consider the following greedy procedure for constructing $N_\epsilon$ (which we don't actually need to implement – it's just for the proof argument):

- Set $N_\epsilon = \{\}$
- While such a point exists, choose an arbitrary point $v \in S_{\mathcal{U}}$ where $\nexists x \in N_\epsilon$ with $\|v - x\| \leq \epsilon$. Set $N_\epsilon = N_\epsilon \cup \{v\}$.

After running this procedure, we have $N_\epsilon = \{x_1, \ldots, x_{|N_\epsilon|}\}$ points that satisfy the condition $\min_{x \in N_\epsilon} \|v - x\| \leq \epsilon$ for all $v \in S_{\mathcal{U}}$. So we just need to bound $|N_\epsilon|$.

To do so, we note that, for all $i, j$, $\|x_i - x_j\| \geq \epsilon$. If not, then either $x_i$ or $x_j$ would not have been added to $N_\epsilon$ by our greedy procedure. Accordingly, if we place balls of radius $\epsilon/2$ around each $x_i$:

$$B(x_1, \epsilon/2), \ldots, B(x_{|N_\epsilon|}, \epsilon/2)$$

then for all $i, j$, $B(x_i, \epsilon/2)$ does not intersect $B(x_j, \epsilon/2)$.

The volume of a $d$ dimensional ball of radius $r$ is $cr^d$ for some value $c$ that does not depend on $r$. So the total volume of $B(x_1, \epsilon/2) \cup \ldots \cup B(x_{|N_\epsilon|}, \epsilon/2)$ is $|N_\epsilon| \cdot c\left(\frac{\epsilon}{2}\right)^d$. At the same time, $B(x_1, \epsilon/2), \ldots, B(x_{|N_\epsilon|}, \epsilon/2)$ are contained inside a ball of radius $1 + \epsilon/2$, which has volume $< c2^d$. So we have:

$$|N_\epsilon| \cdot c\left(\frac{\epsilon}{2}\right)^d < 2^d \qquad \text{which implies} \qquad |N_\epsilon| \leq \left(\frac{4}{\epsilon}\right)^d.$$

□

### Extension to all vectors.

We are now ready to prove Theorem 8.

*Proof.* Choose $m = O\left(\frac{\log(|N_\epsilon|/\delta)}{\epsilon^2}\right) = O\left(\frac{d\log(1/\epsilon)+\log(1/\delta)}{\epsilon^2}\right)$ so that (9) holds for all $x \in N_\epsilon$.

Now consider any $v \in S_{\mathcal{U}}$. It's not hard to see that, for some $x_0, x_1, x_2 \ldots \in N_\epsilon$, $v$ can be written:[8]

$$v = x_0 + c_1 x_1 + c_2 x_2 + \ldots$$

for constants $c_1, c_2, \ldots$ where $|c_i| \leq \epsilon^i$. Applying triangle inequality, we have

$$\begin{aligned}
\|\Pi v\|_2 &= \|\Pi x_0 + c_1 \Pi x_1 + c_2 \Pi x_2 + \ldots\|_2 \\
&\leq \|\Pi x_0\| + \epsilon \|\Pi x_1\| + \epsilon^2 \|\Pi x_2\|_2 + \ldots \\
&\leq (1+\epsilon) + \epsilon(1+\epsilon) + \epsilon^2(1+\epsilon) + \ldots \\
&\leq 1 + O(\epsilon).
\end{aligned}$$

Similarly,

$$\begin{aligned}
\|\Pi v\|_2 &= \|\Pi x_0 + c_1 \Pi x_1 + c_2 \Pi x_2 + \ldots\|_2 \\
&\geq \|\Pi x_0\| - \epsilon \|\Pi x_1\| - \epsilon^2 \|\Pi x_2\|_2 - \ldots \\
&\geq (1-\epsilon) - \epsilon(1+\epsilon) - \epsilon^2(1+\epsilon) - \ldots \\
&\geq 1 - O(\epsilon).
\end{aligned}$$

So we have proven

$$1 - O(\epsilon) \leq \|\Pi v\|_2 \leq 1 + O(\epsilon)$$

for all $v$ in $S_{\mathcal{U}}$. As discussed early, this is sufficient to prove the theorem. □

## References


[1] William B. Johnson and Joram Lindenstrauss. Extensions of Lipschitz mappings into a Hilbert space. Contemporary Mathematics, 1984.

[2] Noga Alon. Problems and results in extremal combinatorics–I. Discrete Mathematics, 273(1-3):31– 53, 2003.

[3] Kasper Green Larsen and Jelani Nelson. Optimality of the Johnson-Lindenstrauss lemma. FOCS, 2017.


[8] To see this, observe that there is some point $x_0$ within distance $\varepsilon$. This will be our first $x_0$. Now, we just need to write the point $v - x_0$, which has norm at most $\varepsilon$. So instead, we could write the point $(v - x_0)/||v - x_0||_2$, which has norm one, and multiply the resulting coefficients by $\varepsilon$. Again there is some point $x_1$ within distance $\varepsilon$, which we take as our next point. We repeat this procedure ad infinitum, each time resulting in a partial sum that gets exponentially closer to $v$.