COMS W6998-5: Algorithms through Geometric Lens (Fall '17) Sep 20, 2017

Lecture 5 – Johnson-Lindenstrauss lemma, Measure Concentration

Instructors: *Alex Andoni, Ilya Razenshteyn* Scribe: *Sandip Sinha*

In this lecture, we will state the most important and famous result in the field of dimension reduction - the Johson-Lindenstrauss lemma, and set up the machinery required to prove it. In particular, we will state several measure concentration results, which are much stronger than the Chebyshev inequality. These results upper bound the probability that the mean (equivalently, sum) of a set of random variables deviate significantly from its expectation.

1 Johnson-Lindenstrauss Lemma

Suppose we are given a set X of n points in high-dimensional space $\mathbb{R}^d$, where d is very large. We would like to map these points to a space of dimension $d' << d$ such that the "geometry" of the set X is (approximately) preserved.

While this is of independent mathematical interest, it also has huge importance for algorithm design. If such a mapping is possible, there are two immediate benefits for algorithms that operates on vectors - space reduction and time reduction. First, one needs to store descriptions of low-dimension vectors, which reduces the space requirement significantly. Moreover, algorithms can operate on the low-dimensional vectors and return approximate answers much faster than it would if it operated on the original vectors.

There are be various notions of geometry of the space that is relevant to a particular problem. In this lecture and the next, we will be interested in approximately preserving Euclidean (ℓ_2) distances between all pairs of points in X. Formally, given $X \subset \mathbb{R}^d, |X| = n$ and fixed $\epsilon > 0$, we seek the existence of a map

$$f : X \to \mathbb{R}^{d'}$$

for some $d' << d$, such that

$$\forall x_1, x_2 \in X, \|f(x_1) - f(x_2)\|_2 \in (1 \pm \epsilon)\|x_1 - x_2\|_2 \qquad (*)$$

Further, we will be interested in the trade-off between the parameters n, d, d' and ϵ.

Theorem 1. *Johnson-Lindenstrauss Lemma [1984]*
Fix $\epsilon > 0$, and let $X \subset \mathbb{R}^d, |X| = n$. Then there exists a function $f : X \to \mathbb{R}^{d'}$ s.t. (*) holds, where $d' = O\left(\frac{\log n}{\epsilon^2}\right)$.

An important feature of this result is that the target dimension d' does not depend on the original dimension d, and depends mildly on the number of points n. This result was shown to be tight recently by Larsen and Nelson [1].

Theorem 2. *[Larsen, Nelson, FOCS 2017]*
For any integers $d, n \geq 2$ and $1/(min\{n, d\})^{0.4999} < \epsilon < 1$, there exists a set of n vectors $X \subset \mathbb{R}^d$ such

that any embedding $f : X \to \mathbb{R}^{d'}$ satisfying (*) must have

$$d' = \Omega\left(\frac{\log n}{\epsilon^2}\right).$$

The proof of the JL Lemma will involve a technique known as random projections. To analyze the quality of these random projections, we will need strong measure concentration results, so the rest of this lecture will be devoted to this aspect.

2 Measure Concentration

To state and compare the measure concentration results, we will use a simple running example. Consider a random walk on a line, where at time 0 we start at the origin 0, and at each time point $i \in \mathbb{N}$, we move 1 unit left or right with equal probability (independent of all other steps). We are interested in the final position after n steps.

Note: For this lecture, we will use the notation $X \in_r D$ to denote that X is drawn from the uniform distribution on D.

Formally, for each $i \in \mathbb{N}$, let $X_i \in_r \{-1, 1\}$ be the random variable which records the step at time i. Let $S_n = X_1 + X_2 + \cdots + X_n$. Then S_n is an integer-valued random variable which represents the final position after n steps. Now, S_n is a sum of independent and identically distributed (iid) binary random variables X_i. By definition, $\mathbb{E}[X_i] = 0$, so linearity of expectation implies $\mathbb{E}[S_n] = 0$.

We are interested in results of the following form:
Meta-statement: $\Pr[S_n$ deviates a lot from $\mathbb{E}[S_n]]$ is small, where S_n is a sum of identically distributed random variables.

2.1 Chebyshev's inequality

The first such result is Chebyshev's inequality, which applies to random variables with finite variance:

$$\Pr[|S_n - \mathbb{E}[S_n]| \geq t\sqrt{\mathrm{Var}[S_n]}] \leq \frac{1}{t^2} \forall t \geq 1.$$

In our case, $\mathrm{Var}[X_i] = 1$. As X_i's are independent (in particular, pairwise independent),

$$\mathrm{Var}[S_n] = \sum_{i=1}^{n} \mathrm{Var}[X_i] = n.$$

Applying Chebyshev's inequality for our example, we get

$$\Pr[|S_n| \geq t\sqrt{n}] \leq \frac{1}{t^2} \ \forall t \geq 1.$$
$$\Pr[|S_n| \geq 3\sqrt{n}] \leq 0.12.$$

2.2 Chernoff bound

The next, much stronger bound that we can use is Chernoff bound, which applies to sums of iid binary random variables. In our case, this bound gives:

$$\Pr[|S_n| \geq t\sqrt{n}] \leq 2\exp\left(-\frac{t^2}{2}\right).$$
$$\Pr[|S_n| \geq 3\sqrt{n}] \leq 0.023.$$

Clearly, Chernoff bound is much stronger. This can be seen from the actual number for the threshold $3\sqrt{n}$, and the rate of decay of the function in terms of t (inverse exponential, which is much faster than inverse polynomial). However, Chernoff bound only applies to sums of iid binary random variables, whereas Chebyshev inequality can be used for any random variable with finite variance.

We note that the Chernoff bound can be improved slightly. This brings us to the Central Limit Theorem (CLT).

2.3 Central Limit Theorem

Theorem 3. *CLT*
Let $X_i, i \in [n]$ be iid random variables with $Var[X_i] < \infty$. Define $S_n := \sum_{i=1}^{n} X_i$. Then, as $n \to \infty$,

$$\frac{S_n - \mathbb{E}[S_n]}{\sqrt{Var[S_n]}} \xrightarrow{d} N(0,1).$$

Here $N(0,1)$ is the standard Gaussian distribution with mean 0 and variance 1. If $X \sim N(0,1)$ then X has probability density function (pdf)

$$f_{N(0,1)}(t) = \frac{1}{\sqrt{2\pi}}\exp\left(-\frac{t^2}{2}\right).$$

We specify the sense of convergence in CLT. We say that a sequence of random variables $\{X_n\}$ converges to a random variable X *in distribution* and denote it by $X_n \xrightarrow{d} X$ if the cumulative density function (CDF) $F_{X_n}(t)$ of X_n converges to the CDF $F_X(t)$ of X point-wise for all $t \in \mathbb{R}$. In our case, this means

$$\Pr\left[\frac{S_n}{\sqrt{n}} \geq t\right] \xrightarrow{n\to\infty} \Pr_{X\sim N(0,1)}[X \geq t] = \frac{1}{\sqrt{2\pi}}\int_t^\infty \exp\left(-\frac{x^2}{2}\right)dx \text{ for all } t \in \mathbb{R}$$

This is the precise convergence guarantee given by CLT. We use the upper bound

$$\frac{1}{\sqrt{2\pi}}\int_t^\infty \exp\left(-\frac{x^2}{2}\right)dx \leq \frac{1}{\sqrt{2\pi}t}\exp\left(-\frac{t^2}{2}\right) \text{ for all } t > 0$$
$$\text{So, } \lim_{n\to\infty}\Pr[|S_n| \geq t\sqrt{n}] \leq \sqrt{\frac{2}{\pi}}\frac{1}{t}\exp\left(-\frac{t^2}{2}\right)$$
$$\lim_{n\to\infty}\Pr[|S_n| \geq 3\sqrt{n}] \leq 0.002$$

While this is a slight improvement over the Chernoff bound, it holds only in the limit, and so it cannot be applied for a fixed finite n, unlike Chernoff bound. However, it can be applied to sums of iid random variables with finite variance, whereas Chernoff bound only applies to sums of iid binary random variables.

Another way to look at this mode of convergence guaranteed by CLT is to look at the histogram of

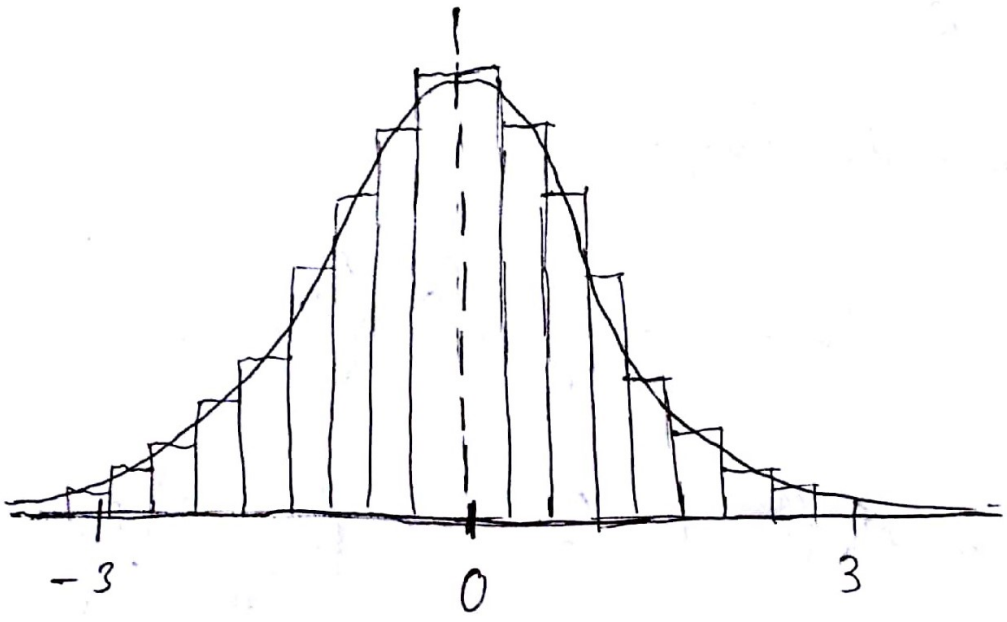

Figure 1: The histogram of the normalized sum approaches the pdf of $N(0,1)$ distribution as $n \to \infty$.

the normalized sum of random variables. The histogram of this (discrete, in our case) random variable approaches the pdf of a normal distribution as $n \to \infty$, as shown in Figure 1.

2.4 Uniform distribution over Boolean Hypercube

In our example, we had n iid binary random variables $X_i \in_r \{-1, 1\}$. Associating each of them with a coordinate, we can think of each realization of these points $X_i = x_i$ as a point $x = (x_1, \cdots, x_n) \in \{-1, 1\}^n$. We call $\{-1, 1\}^n$ the Boolean Hypercube, a generalization of a square ($n = 2$) or cube ($n = 3$) to n dimensions. Then it is easy to see that a uniformly random point from this hypercube can be drawn by drawing each $X_i \in_r \{-1, 1\}$ uniformly and independently, as we have done in our example. Then we have

$$S_n = \sum_{i=1}^{n} X_i = 2 \cdot \#\{i \in [n] : X_i = 1\} - n.$$

The previous results tell us that under the uniform distribution on the hypercube, most (at least 99%) of the mass is concentrated on the vectors $x \in \{-1, 1\}^n$ such that the number of 1's in x is $\frac{n}{2} \pm O(\sqrt{n})$, as shown in Figure 2.

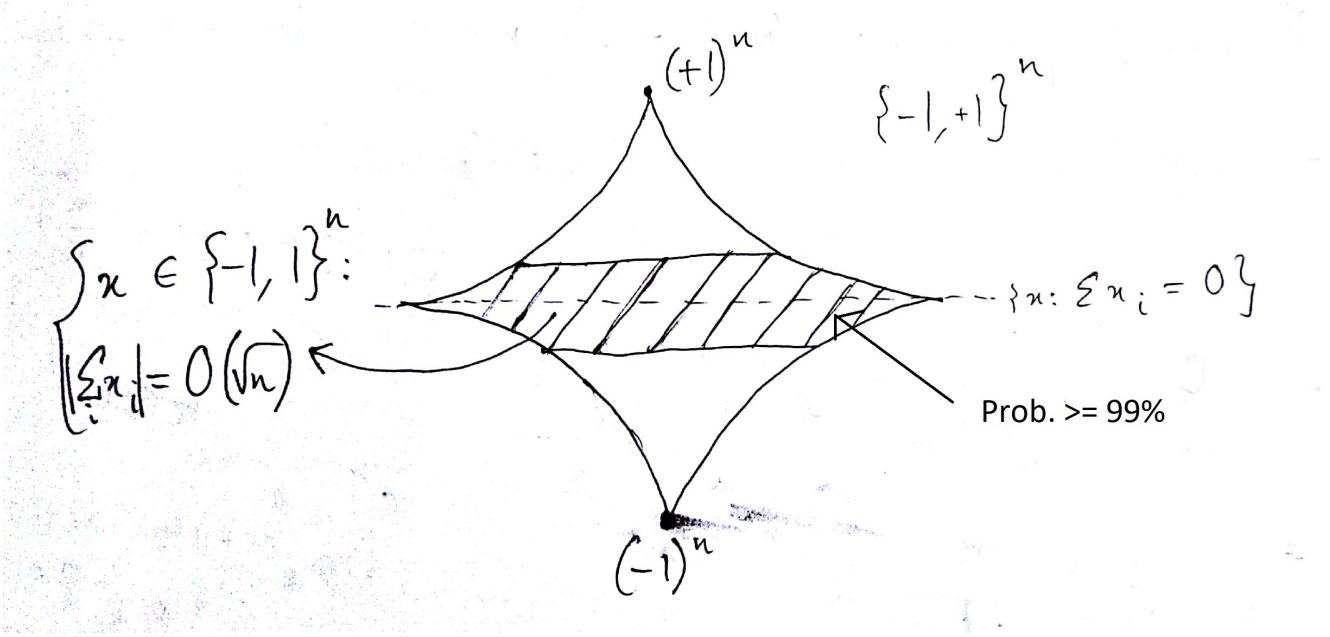

Figure 2: The Boolean Hypercube $\{-1,1\}^n$. The width at a particular level represents the number of sequences x with $\sum_i x_i = c$ for $c \in \{-n, -n+1, \cdots, n\}$.

2.5 Uniform distribution over the Hypersphere

Now, we would like to characterize the uniform distribution over the unit hypersphere, defined below:

$$S^{d-1} := \{x \in \mathbb{R}^d | \|x\|_2 = 1\}.$$

Note that we write S^{d-1} to denote the hypersphere of points in $\mathbb{R}^d$ because it is a $d-1$-dimensional surface. Equivalently, it can be completely described by $d-1$ parameters. This is easy to see if we switch to polar coordinates. For instance, if $d = 2$, we can uniquely identify a point p on the unit circle using only 1 parameter: the angle $\theta \in (-\pi, \pi]$ that x makes with the positive x-axis.

By definition, all points on the unit hypersphere have the same ℓ_2-norm. This norm is invariant under rotations. So, under the uniform distribution, the mass it assigns to a section of the sphere must remain the same if it is rotated. We formalize the idea of rotation as being an orthogonal matrix, i.e., a square real matrix $U : \mathbb{R}^d \to \mathbb{R}^d$ such that $U^T U = UU^T = \mathbf{I}$. For such a matrix U, we have, for any vector x,

$$\|Ux\|_2^2 = (Ux)^T(Ux) = x^T U^T U x = x^T x = \|x\|_2^2,$$

so U preserves ℓ_2 norm. This leads us to the first (indirect) definition of the uniform distribution.

Definition 4. *Fix $d \in \mathbb{N}$. The uniform distribution on S^{d-1} is the unique distribution D such that for all orthogonal matrices $U \in \mathbb{R}^{d \times d}$ and all measurable sets $A \subset S^{d-1}$, the following holds:*

$$\Pr_{x \sim D}[x \in A] = \Pr_{x \sim D}[Ux \in A]$$

While we do not give the precise definition of a set being measurable, we remark that it intuitively

means that the d-dimensional volume of the set is well-defined.

Clearly, this definition does not tell us how to sample from the uniform distribution. A simple calculation shows that even for $d = 2$, if we draw $x_1, x_2 \in_r [-1, 1]$ (from the uniform distribution on $[-1, 1]$) independently and normalize $x = (x_1, x_2)$ by $\|x\|_2$, this does not give the uniform distribution on the circle. Now, we outline the correct procedure:

- Sample $g_1, g_2, \cdots, g_d \sim N(0, 1)$ independently. Let $g = (g_1, g_2, \cdots, g_d) \in \mathbb{R}^d$.
- Return $g' = \frac{g}{\|g\|_2}$.

Claim 5. *Fix $d \in \mathbb{N}$. Let g' be as defined above. Then the distribution of g' is uniform over S^{d-1}.*

To prove this claim, we need another claim:

Claim 6. *Fix $d \in \mathbb{N}$. Let g be as defined above. Then the distribution of g is spherically symmetrical. In other words, if f_g is the pdf of this distribution, then $f_g(u)$ depends only on $\|u\|_2$.*

Proof. Let $g = (g_1, \cdots, g_d)$. As g_i are independent, the density of g is the product of the densities of each g_i. For a fixed vector $u = (u_1, u_2, \cdots, u_n)$, the density $f_g(u)$ is given by

$$f_g(u) = \prod_{i=1}^{d} f_{N(0,1)}(u_i) = \frac{1}{(2\pi)^{d/2}} \prod_i \exp\left(-\frac{u_i^2}{2}\right) = \frac{1}{(2\pi)^{d/2}} \exp\left(-\frac{\sum_i u_i^2}{2}\right) = \frac{1}{(2\pi)^{d/2}} \exp\left(-\frac{\|u\|_2^2}{2}\right).$$

□

This Claim immediately implies Claim 5, since all points on the sphere have equal (unit) ℓ_2 norm and hence have equal density.

Now, suppose $x = (x_1, \cdots, x_d) \in_r S^{d-1}$. We want to understand the distribution of x_1. By symmetry, $\mathbb{E}[x_1] = 0$. Let $x = \frac{g}{\|g\|_2}$, where $g = (g_1, \cdots, g_d)$ is generated by drawing each $g_i \sim N(0, 1)$ iid. We have $\mathbb{E}[g_i^2] = \mathrm{Var}[g_i] = 1$. So, by abuse of notation, we can write

$$x_1 = \frac{g_1}{\sqrt{g_1^2 + \cdots + g_d^2}} \approx \frac{N(0, 1)}{\sqrt{d}}$$

by which we mean that we expect x_1 to behave as a standard normal, scaled by $1/\sqrt{d}$.

Definition 7. *The chi-squared distribution with d degrees of freedom is the random variable defined as*

$$\chi^2(d) := g_1^2 + g_2^2 + \cdots + g_d^2$$

where $g_1, g_2, \cdots, g_d \sim N(0, 1)$ are iid random variables.

Clearly, $\mathbb{E}[\chi^2(d)] = d$. By CLT, we expect $\chi^2(d)$ to concentrate in a small window around d. Equivalently, we expect $\sqrt{\sum_i g_i^2}$ to concentrate in a small window around $\sqrt{d}$, which implies that $\Pr\left[|x_1| \geq \frac{3}{\sqrt{d}}\right]$ is small. Moreover, this is true in all directions (not just coordinate directions) by rotational symmetry. Let $u \in S^{d-1}$ be a fixed vector, and let c be some constant (say $c = 3$). Define the equator w.r.t u by

$$E_c(u) := \left\{x \in S^{d-1} : |\langle x, u\rangle| \leq \frac{c}{\sqrt{d}}\right\}.$$

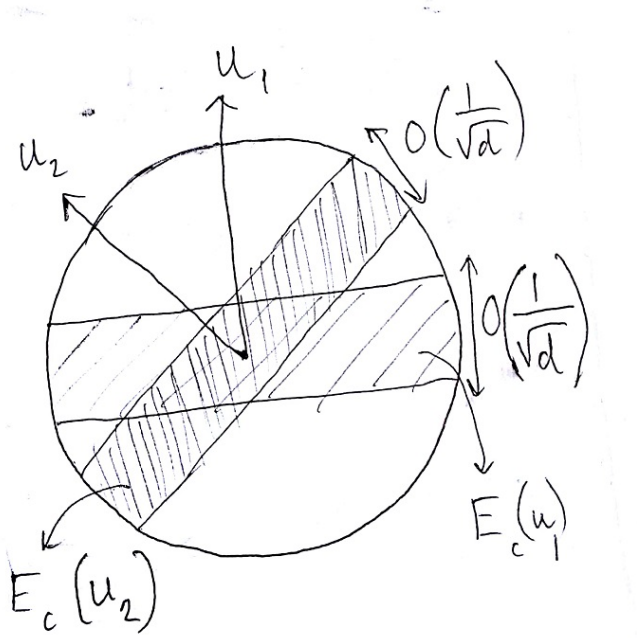

Figure 3: The hypersphere S^{d-1} with the equators $E_c(u)$ corresponding to 2 vectors u_1 and u_2 highlighted. Each of the shaded regions contains most of the mass under the uniform distribution.

Then $\Pr_{x \in_r S^{d-1}}[x \notin E_c(u)]$ is very small. This is true for all $u \in S^{d-1}$ simultaneously, as shown in Figure 3. So, for any fixed vector $u \in S^{d-1}$, most points on the sphere are nearly orthogonal to it.

The following lemma says that we there exists a set containing lots of points in the hypersphere such that any two points in the set are nearly orthogonal to each other.

Lemma 8. *Fix $m \in \mathbb{N}, \epsilon > 0$. Then there exists $x_1, x_2, \cdots, x_n \in S^{m-1}$, where $n = 2^{\Omega(\epsilon^2 m)}$, such that*

$$\forall i \neq j,\ |\langle x_i, x_j \rangle| \leq \epsilon.$$

The proof is via the probabilistic method. This method is useful when want to show the existence of a "good" object. To show this existence, it is enough to define a probability distribution over all objects in a relevant space, and show that the probability that an object sampled according to this distribution is good is strictly positive. While this may seem a very weak implication, it often leads to a much simpler proof than if one were to give an explicit construction. Indeed, in many cases, this is the only method known to show existence of certain objects. We remark that in order to extract efficient algorithms from this method, we need a stronger guarantee: for instance, the probability of an object being "good" is at least a constant or $\frac{1}{\text{poly}(n)}$, where n is the input size, so that a randomized algorithm can draw polynomially many samples and check if they are good, and succeed with high probability.

Proof. Fix $i \neq j$. Let $x_i, x_j \in_r S^{m-1}$. We can apply the same rotational transformation U to both x_i and x_j such that $Ux_i = e_1$ and $Ux_j = \tilde{x}$, where $e_1 = (1, 0, 0, \cdots, 0)$ is the first coordinate vector and $\tilde{x}$ is some unit vector. By rotational invariance of the uniform distribution, we have that $\tilde{x}$ is also uniformly

distributed over S^{m-1}.

$$\langle x_i, x_j \rangle = \langle e_1, \tilde{x} \rangle = \tilde{x}_1 \sim \frac{N(0,1)}{\sqrt{m}}$$

By the earlier arguments using CLT, we have

$$\Pr[|\langle x_i, x_j \rangle| \geq \epsilon] \leq \exp\left(-\Theta(\epsilon^2 m)\right)$$

By the union bound over $\binom{n}{2} \leq n^2$ pairs where n is as stated in the lemma, we have that

$$\Pr[\exists i \neq j \text{ s.t. } |\langle x_i, x_j \rangle| \geq \epsilon] \leq n^2 \exp\left(-\Theta(\epsilon^2 m)\right) << 1.$$

So, with positive probability, the n random vectors are pairwise almost orthogonal, which implies there exists a set of n vectors with the same property. □

We will end this lecture with an application of this lemma to prove the JL Lemma in a special case. Fix $d \in \mathbb{N}$. We define a simplex as a set of points Δ_d of $d+1$ points in $\mathbb{R}^d$ such that any two points are at distance 1 from each other

Lemma 9. *Johnson-Lindenstrauss Lemma for a simplex Let Δ_d be a simplex in $\mathbb{R}^d$. Then there exists a map $f : \Delta_d \to \mathbb{R}^m$ with $m = O\left(\frac{\log |\Delta_d|}{\epsilon^2}\right)$ such that (*) holds.*

Proof. Let $S = \{x_1, \cdots, x_n\}$ be the points in S^{m-1} guaranteed by Lemma 8. Define f by mapping the points in Δ_d to points in S such that the function is injective (one-one). Fix $i \neq j$.

$$\|x_i - x_j\|_2^2 = \|x_i\|_2^2 + \|x_j\|_2^2 - 2\langle x_i, x_j \rangle \in 2(1 \pm \epsilon)$$

So, the pairwise distances in the target space satisfy $\|x_i - x_j\|_2 \in \sqrt{2}(1 \pm \epsilon)$. We can scale f by $\sqrt{2}$ in each coordinate and get the guarantee in the JL Lemma. Now, we have to ensure that the number of points n in the target space is at least $d+1$, so that f is one-one.

$$n = 2^{\Omega(\epsilon^2 m)} \geq d+1$$

So we need to set $m \geq \Theta\left(\frac{\log(d+1)}{\epsilon^2}\right) = \Theta\left(\frac{\log |\Delta_d|}{\epsilon^2}\right)$. Thus, for $m = O\left(\frac{\log |\Delta_d|}{\epsilon^2}\right)$, the lemma holds. □

References

[1] Kasper Green Larsen and Jelani Nelson. Optimality of the johnson-lindenstrauss lemma. *CoRR*, abs/1609.02094, 2016.