CS-621 Theory Gems November 28, 2012

Lecture 21

Lecturer: Aleksander Mądry *Scribes: Alhussein Fawzi, Dorina Thanou*

1 Introduction

Today, we will briefly discuss an important technique in probability theory – *measure concentration*. Roughly speaking, measure concentration corresponds to exploiting a phenomenon that some functions of random variables are highly concentrated around their expectation/median. The main example that will be of our interest here is Johnson-Lindenstrauss (JL) lemma. The JL lemma is a very powerful tool for dimensionality reduction in high-dimensional Euclidean spaces and it is widely used to alleviate the curse of dimensionality that occurs in applications where one needs to deal with high-dimensional data.

2 Examples of Measure Concentration

Probably the most well-known example of measure concentration result states that the sum of independent random variables is tightly concentrated around its expectation/median. In particular, if $X_1, X_2, ..., X_n$ are independent and identically distributed random variables (i.i.d.) with each X_i taking a value $X_i \in \{1, -1\}$ with equal probability, the celebrated Chernoff bound states that their sum $X = \sum_{i=1}^{n} X_i$ is highly concentrated around its expectation. Specifically, the probability that $|X| > t$ is exponentially decaying with t, i.e.,

$$Pr(|X| > t) < 2e^{-\frac{t^2}{2n}}. \tag{1}$$

(Note that expectation of X is just zero.)

Although this result is the most well-known one and it already has plethora of applications, it can actually be seen as a special case of a more general measure concentration phenomena.

To this end, let us focus our attention on general real functions on hypercube and say that a function $f : \{-1, 1\}^n \to \mathbb{R}$ is *L-Lipschitz*, for some $L > 0$, (with respect to ℓ_1 metric) iff, for all $x, y \in \{-1, 1\}^n$,

$$|f(x) - f(y)| \leq L\|x - y\|_1. \tag{2}$$

(One can view the L-Lipschitz condition as a quantified version of uniform continuity of f.)

Now, one can show that for *any* 1-Lipschitz function of n random variables $X_1, X_2, \ldots, X_n$ that are i.i.d. and are $+1$ and -1 with equal probability, an analogous to (1) concentration around the *median* μ of f occurs. Namely, we have

$$Pr(f(X_1, \ldots, X_n) > \mu + t) < 2e^{-\frac{t^2}{2n}}. \tag{3}$$

(One can get a result for arbitrary Lipschitz constant L just by scaling.)

As the sum function is clearly 1-Lipschitz, one can see that Chernoff bound is indeed a consequence of this more general statement.

3 The Johnson-Lindenstrauss Lemma

The main example of measure concentration phenomenon that we want to focus on today is captured by *Johnson-Lindenstrauss (JL) lemma* and corresponds to the behavior of random vectors on a high-dimensional unit sphere.

Roughly speaking, the Johnson-Lindenstrauss lemma tells us that the ℓ_2-distance of high-dimensional vectors is well preserved under *random* projection to a (much) lower dimension.

Lemma 1 (Johnson-Lindenstrauss lemma) *Consider a set of n vectors $x^i \in \mathbb{R}^d$ and a random k-dimensional subspace of $\mathbb{R}^d$. Let y^i be the projection of each x^i on that subspace. For any $\varepsilon > 0$, if $k = \Omega(\epsilon^{-2} \log n)$ then with probability at least $1 - \frac{1}{n}$,*

$$(1-\epsilon)\sqrt{\frac{k}{d}}\|x^i - x^j\|_2 \leq \|y^i - y^j\|_2 \leq (1+\epsilon)\sqrt{\frac{k}{d}}\|x^i - x^j\|_2, \quad \forall i, j. \tag{4}$$

In the light of this lemma, if we have some high-dimensional data whose key characteristic we are interested in is captured by ℓ_2-distance, then we can achieve even an exponential compression of this data's dimension at the price of introducing only $(1+\varepsilon)$ error (note that the $\sqrt{\frac{k}{d}}$ is just a normalizing scaling factor).

It turns out that there is a lot of scenarios (especially, in statistics and machine learning) where this technique is applicable and allows one to lift the "curse of dimensionality". Namely, in a lot of applications, (very) high-dimensional data arises naturally and this kind of compression – often called *dimensionality reduction* – provides a powerful tool for dealing with the computational cost of processing such data.

3.1 Random Subspaces

Before proceeding to the proof of this lemma, we first need to make the notion of random subspace precise.

To this end, let us start by defining what we mean by a *random unit vector* $x \in \mathbb{S}^{d-1}$, where $\mathbb{S}^{d-1}$ is the d-dimensional unit sphere. We will view such a vector as a result of a generation procedure in which, first, we sample each of its d coordinates independently from a Gaussian distribution $\mathcal{N}(0,1)$ that has zero mean and standard deviation one; and then normalize it to make its norm equal to 1. (Note that one of the important and desirable properties of this definition is that the resulting probability measure on the sphere is rotationally invariant.)

Once we defined our notion of random unit vector, i.e., we defined our probability measure on the sphere, we can proceed to defining what we mean by a *random subspace* of dimension k. Again, we will do this by specifying the random process that generates it. This process is as follows:

1. Choose a random unit vector and make it the first basis vector v^1 of the subspace.
2. For the next $k-1$ rounds repeat the following: pick a random unit vector, subtract from it its projection on the subspace spanned by the previously chosen vectors $v^1, \ldots, v^{i-1}$, and normalize it to form the next basis vector v^i.[1]

Clearly, after this procedure is finished we end up with an orthonormal basis $v^1, \ldots, v^k$ that spans the desired (random) subspace of dimension k. (Note that the above procedure is nothing else than Gram-Schmidt orthogonalization applied to a set of random unit vectors $\{v^1, \ldots, v^k\}$.)

Also, one can see that under this definition the projection y^i of a data point x^i onto such a random subspace can be written in a matrix form as

$$\underbrace{\begin{bmatrix} v_1^1 & v_2^1 & \ldots & v_d^1 \\ v_1^2 & v_2^2 & \ldots & v_d^2 \\ \vdots & & & \\ v_1^k & v_2^k & \ldots & v_d^k \end{bmatrix}}_{V} \begin{bmatrix} x_1^i \\ x_2^i \\ \vdots \\ \vdots \\ x_d^i \end{bmatrix} = \begin{bmatrix} y_1^i \\ y_2^i \\ \vdots \\ y_k^i \end{bmatrix}.$$

Here, each row of the *projection matrix* V corresponds to one random basis vector v^i.

[1] It is easy to see that this randomly chosen unit vector is not in the span of the vectors $v^1, \ldots, v^{i-1}$ with probability 1.

3.2 Proof of the JL Lemma

Now that we have defined what a random vector and what a random subspace is, we are ready to prove the Johnson-Lindenstrauss lemma. As a first step, we show that this lemma follows from a simpler statement that just focuses on the norm of the projection of a *fixed* vector x in d dimensions onto a random k-dimensional subspace.

Lemma 2 *Let x be an arbitrary vector in $\mathbb{R}^d$ and $z \in \mathbb{R}^k$ be its projection onto a random k-dimensional subspace. Then, for any $\varepsilon > 0$, as long as $k = \Omega(\epsilon^{-2} \log n)$, we have*

$$\left| \frac{\|z\|_2}{\|x\|_2} - \sqrt{\frac{k}{d}} \right| \leq \epsilon \sqrt{\frac{k}{d}},$$

with probability exceeding $1 - \frac{1}{n^3}$.

It is not hard to see that once we prove Lemma 2, the Johnson-Lindenstrauss lemma follows easily. Indeed, by applying the above lemma with $x = x^i - x^j$, for any fixed i and j, we get

$$Pr\left[\left| \frac{\|z^{i,j}\|_2}{\|x^i - x^j\|_2} - \sqrt{\frac{k}{d}} \right| \leq \epsilon \sqrt{\frac{k}{d}} \right] \geq 1 - \frac{1}{n^3},$$

where $z^{i,j}$ is the projection of $x^i - x^j$ on the random subspace. Since the projection is a linear map, we have $z^{i,j} = y^i - y^j$. So, applying a union bound to the previous inequality, over all $O(n^2)$ pairs (i, j), we get that

$$Pr\left[\forall i \neq j, \left| \frac{\|y^i - y^j\|_2}{\|x^i - x^j\|_2} - \sqrt{\frac{k}{d}} \right| \leq \epsilon \sqrt{\frac{k}{d}} \right] \geq 1 - \frac{n(n-1)}{2n^3} \geq 1 - \frac{1}{n},$$

which can be easily seen to be equivalent to the statement of Johnson-Lindenstrauss lemma.

Hence, from now on we focus on proving Lemma 2. (Observe that by scaling, it suffices to prove this lemma for the case of x being a unit vector.) To make our task easier, we want to first invert our perspective. Namely, instead of looking at the norm of a projection of an arbitrary vector onto a random k-dimensional subspace, we prefer to look at the norm of a projection of a *random* vector on a *fixed* k-dimensional subspace that corresponds to the first k coordinates of that vector.

It is not hard to see that these two views are completely equivalent. To this end, note that we can always rotate the space in such a way that the random k-dimensional subspace we chosen is just the projection onto the first k coordinates. Formally, let U denote the $d \times d$ unitary matrix whose first k rows are equal to vectors v^is (that form the basis of the random subspace we have chosen) and where the remaining rows are chosen arbitrarily to form an orthonormal basis of the orthogonal complement of our subspace. Then, we have that

$$z_i = (v^i)^T x = (U^{-1} v^i)^T (U^{-1} x),$$

for any $1 \leq i \leq k$, as $U^{-1} = U^T$ is a unitary matrix too and thus satisfies $(U^{-1})^T U^{-1} = I$. Since $U^{-1} v^i$ is equal to the i-th standard basis vector e^i and $U^{-1} x$ is a random vector (as it corresponds to a random rotation of a fixed vector), it is indeed valid to see z as the projection of a random vector onto the subspace spanned by its first k coordinates.

Thanks to the above simplification of the perspective, our goal now is to study how the norm of the first k coordinates of a random vector (of unit norm) concentrates around a particular value. To this end, note that if $z' = (z_1, \ldots, z_d) = (z, z_{k+1}, \ldots, z_d)$ is a random unit vector then clearly we have

$$\mathbb{E}\left(\sum_{i=1}^{d} z_i^2 \right) = 1.$$

Since each z_i's are identically distributed, we obtain

$$\mathbb{E}\left(\sum_{i=1}^{k} z_i^2\right) = \frac{k}{d}.$$

Thus indeed ℓ_2^2-norm of the k first coordinates of a random vector has the desired expectation. However, to prove Lemma 2, we also need to study how this norm is concentrated around its expectation.

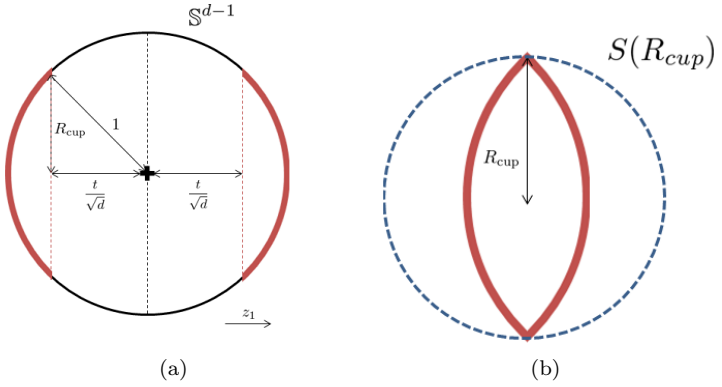
We will not do this today. Instead, just to give a flavor of involved techniques, we prove here a simpler result that bounds the concentration of the corresponding norm for $k = 1$. Specifically, we show that the probability that $|z_1|$ is larger than $\frac{t}{\sqrt{d}}$ is exponentially decaying with t.

Lemma 3 *Let $z' = (z_1, \ldots, z_d)$ be a random vector in $\mathbb{S}^{d-1}$. We have*

$$Pr\left(|z_1| > \frac{t}{\sqrt{d}}\right) \leq 2\exp\left(-\frac{t^2}{2}\right),$$

for any $0 < t \leq \sqrt{\frac{d}{2}}$.

Proof The proof of this lemma is based on a simple geometric argument. Let us fix some $t > 0$. As z'



(a) (b)

Figure 1: Illustration of the proof in two dimensions. (a) The caps corresponding to $|z_1| > \frac{t}{\sqrt{d}}$ are marked with red color. (b) Pictorial argument justifying upperbounding the area of these two caps by the area of corresponding sphere of the same radius.

is a random vector from a d-dimensional unit sphere $\mathbb{S}^{d-1}$, we can see that the probability of choosing z' with $|z_1| > \frac{t}{\sqrt{d}}$ is exactly the ratio of the area of two d-dimensional caps of radius $R_{\text{cup}} = \sqrt{1 - \frac{t^2}{d}}$ to the total area of a unit d-dimensional sphere. (See Figure 1 (a) that represents the situation in two dimensions, i.e., the case of $d = 2$.)

We can upperbound the area of these two caps by the area of a whole sphere of the same radius (see Figure 1 (b)). As the area of a d-dimensional sphere $S(R)$ of radius R has to be a function of the form $C_d \cdot R^{d-1}$, where C_d is some coefficient depending on d but not on R, we have

$$Pr\left(|z_1| > \frac{t}{\sqrt{d}}\right) \leq \frac{\text{area}(S(R_{\text{cup}}))}{\text{area}(\mathbb{S}^{d-1})} = \frac{C_d \cdot R_{\text{cup}}^{d-1}}{C_d \cdot 1^{d-1}} = \left(1 - \frac{t^2}{d}\right)^{\frac{d-1}{2}}.$$

Using the fact that $(1 - x/n)^n \leq \exp(-x)$, we conclude that

$$Pr\left(|z_1| > \frac{t}{\sqrt{d}}\right) \leq 2\exp(-t^2/2),$$

whenever $0 < t \leq \sqrt{\frac{d}{2}}$, as desired. ■

It is interesting to note that by applying Lemma 3 with $t = \Omega(\sqrt{\log n})$, we get that the probability that $|z_1|$ exceeds $\sqrt{\frac{\log n}{d}}$ is bounded by $\frac{1}{n^{O(1)}}$. This tells us that in high dimensions almost all the vectors on the unit sphere are close to being orthogonal. Indeed, thanks to the rotation invariance of the scalar product, we can always take one of the vectors to have its first coordinate be equal to 1 and have all the remaining coordinates equal to zero. Then, the scalar product of a random unit vector z' with this vector is equal to z_1. In high dimensions, the quantity $\frac{\log n}{d}$ is very small, which gives a very small scalar product with high probability.

Unfortunately, as we already mentioned, the bounds provided by the Lemma 3 are too weak to yield the desired concentration of the norm of the projection of z' on the first k coordinates. Therefore, we state (without proof) a stronger version of Lemma 3 that allows one take advantage of larger values of k.

Lemma 4 *Let $z' = (z_1, \ldots, z_d) = (z, z_{k+1}, \ldots, z_d)$ be a random vector in $\mathbb{S}^{d-1}$. We have*

$$Pr\left(\left|\|z\|_2 - \sqrt{\frac{k}{d}}\right| > t\right) \leq 2e^{-\frac{t^2 d}{2}}.$$

Once we have this lemma, the proof of Lemma 2 is straightforward. We just take $t = \varepsilon\sqrt{\frac{k}{d}}$ and $k = 10\varepsilon^{-2}\ln n$. We then have

$$Pr\left(\left|\|z\|_2 - \sqrt{\frac{k}{d}}\right| > \epsilon\sqrt{\frac{k}{d}}\right) \leq \frac{1}{n^3},$$

which proves Lemma 2, and thus the Johnson-Lindenstrauss lemma.

3.3 Further Discussion

As we presented it here, the JL lemma is not very practical. This is so as our generation of the projection matrix V requires performing Gram-Schmidt orthnormalization that is computationally quite expensive when n is large (which is often the case). To circumvent this issue and make JL lemma more practical, there was a lot of (successful) work on developing much more efficient constructions of the projection matrix V. In these latest constructions, this matrix is generated via a very simple and easy to implement procedure that makes V have only few entries of each column being non-zero. As a result, not only the whole construction is very efficient, but also the resulting matrix V is sparse (i.e., it has only small fraction of entries non-zero), which leads to computations of the projections of the input vectors being very efficient too. All of these advancements made JL lemma a truly practical tool.

Given the usefulness of JL lemma in applications that operate based on ℓ_2-distance, it is natural to wonder if similar results could be achieved for other ℓ_p-distances. Unfortunately, it seems that it is not the case and in fact for some of the distances (e.g., ℓ_1-distance) there are strong lowerbounds on the possible dimension reduction. (Also, it is known that for ℓ_2-distance, the dimension reduction offered by JL lemma is essentially optimal.)