Homework 1 -Solutions / MAT280

Prepared by Shuyang Ling

April 20, 2016

Problem 1 Show that two random vectors in high dimensions are almost orthogonal.

Note: In your theorem you need to formalize what "almost orthogonal" means (what it means will come out of your proof). You first need to select a probability distribution of your choice and apply an appropriate concentration inequality (but keep in mind that if e.g. x and y are Gaussian random vectors, then the entries of the inner product $\langle x, y\rangle$ are no longer Gaussian).

Answer: First note that the angle θ_d between two vectors in $\mathbb{R}^d$ is given characterized by

$$\cos\theta_d = \frac{\langle \boldsymbol{x}, \boldsymbol{y}\rangle}{\|\boldsymbol{x}\|\|\boldsymbol{y}\|} = \frac{\langle \boldsymbol{x}, \boldsymbol{y}\rangle}{d}$$

if $\|\boldsymbol{x}\| = \|\boldsymbol{y}\| = \sqrt{d}$. Two vectors are almost orthogonal if $\theta_d \approx \frac{\pi}{2}$ or $\cos\theta_d \approx 0$.

We consider a few different scenarios to illustrate different proof techniques.

Suppose we have two random vectors, $\boldsymbol{x}$ and $\boldsymbol{y} \in \{-1, 1\}^d$, each entry of which takes ± 1 with equal probability. Note that $\langle \boldsymbol{x}, \boldsymbol{y}\rangle = \sum_{i=1}^d x_i y_i$ is the sum of d i.i.d. random variables where $x_i y_i$ yields the same distribution as x_i and y_i, i.e., $\mathbb{P}(x_i y_i = 1) = \mathbb{P}(x_i y_i = -1) = \frac{1}{2}$. Therefore, $\mathbb{E}(\langle \boldsymbol{x}, \boldsymbol{y}\rangle) = \sum_{i=1}^d \mathbb{E}(x_i y_i) = 0$. The easiest way to see the asymptotic behavior of $\cos\theta_d$ is just by applying the Law of large numbers, which says the sample mean of $\{x_i y_i\}_{i=1}^d$ will converge almost surely to its expectation 0. However, the Law of large numbers won't tell us the rate of convergence.

Note that $\{x_i y_i\}_{i=1}^d$ are bounded i.i.d. random variables with mean 0. Hence, one can use the Chernoff-Hoeffding inequality to obtain a quantitative bound. For any given $\epsilon > 0$,

$$\mathbb{P}(|\langle \boldsymbol{x}, \boldsymbol{y}\rangle| \geq d\epsilon) \leq 2\exp\left(-\frac{2d^2\epsilon^2}{d}\right) = 2\exp\left(-2d\epsilon^2\right).$$

In other words, $\mathbb{P}\left(\frac{|\langle \boldsymbol{x}, \boldsymbol{y}\rangle|}{d} \geq \epsilon\right) \leq 2\exp\left(-2d\epsilon^2\right)$, which goes to 0 as the dimension d goes to infinity. The probability of $|\cos\theta_d| \geq \epsilon$ decreases exponentially w.r.t. d. This indicates that two random vectors uniformly drawn from $\{-1, 1\}^d$ tend to be "almost orthogonal" when the dimension is huge. For instance, we can set $\epsilon = \sqrt{\log d}$ and obtain that $\mathbb{P}\left(\frac{|\langle \boldsymbol{x}, \boldsymbol{y}\rangle|}{d} \geq \sqrt{\log d}\right) \leq 2\exp\left(-2d\log d\right)$.

It is slightly more complicated when $\boldsymbol{x}, \boldsymbol{y}$ are both drawn from Gaussian distribution in $\mathbb{R}^d$ since the random variables are no longer bounded. The idea is simple: first truncate the Gaussian random variable, followed by applying Hoeffding's inequality. Simple calculations yield

$$\begin{aligned}
\mathbb{P}(|x_i| \geq t) &= \frac{1}{\sqrt{2\pi}} \int_{|x|\geq t} e^{-\frac{x^2}{2}} dx \leq \frac{1}{\sqrt{2\pi}t} \int_{|x|\geq t} |x| e^{-\frac{x^2}{2}} dx \\
&\leq \frac{1}{t} \int_t^\infty x e^{-\frac{x^2}{2}} dx \leq \frac{1}{t} e^{-\frac{t^2}{2}} \leq e^{-\frac{t^2}{2}}
\end{aligned}$$

for any x_i standard Gaussian random variable and $t > 1$. In this problem, we set $t := \sqrt{2\alpha\log d}$ for some large $\alpha > 1$. Then taking the union bound over all $1 \leq i \leq d$

$$\mathbb{P}(|x_i| \geq t, |y_i| \geq t, \forall 1 \leq i \leq d) \leq 2d^{-\alpha+1}.$$

Let's denote the event $E := \{|x_i| \le t, |y_i| \le t, \forall 1 \le i \le d\}$. In other words, over E, all x_i and y_i are bounded by t with probability at least $1 - 2d^{-\alpha+1}$ and now we can use Hoeffding's inequality. Note that x_i^2, y_i^2 and $|x_i y_i| \le t^2 = 2\alpha \log d$ on E. Applying Hoeffding's inequality and setting $0 < \epsilon < 1$,

$$\mathbb{P}\left(|\sum_{i=1}^{d} x_i y_i| \ge d\epsilon \right) \le 2\exp\left(-\frac{2d^2\epsilon^2}{dt^2}\right) \le 2\exp\left(-\frac{d\epsilon^2}{\alpha \log d}\right)$$

and

$$\mathbb{P}\left(|\sum_{i=1}^{d} x_i^2 - d| \ge d\epsilon \right) = \mathbb{P}\left(|\sum_{i=1}^{d} y_i^2 - d| \ge d\epsilon \right) \le 2\exp\left(-\frac{d\epsilon^2}{\alpha \log d}\right).$$

Therefore, there holds

$$|\sum_{i=1}^{d} x_i y_i| \le d\epsilon, \quad \sum_{i=1}^{d} x_i^2 \ge d(1-\epsilon), \quad \sum_{i=1}^{d} y_i^2 \ge d(1-\epsilon)$$

with probability at least $1 - 6\exp\left(-\frac{d\epsilon^2}{\alpha \log d}\right) - 2d^{-\alpha+1}$ by taking the union bound. Now we have

$$|\cos\theta_d| = \frac{|\langle \boldsymbol{x}, \boldsymbol{y}\rangle|}{\|\boldsymbol{x}\|\|\boldsymbol{y}\|} \le \frac{d\epsilon}{d(1-\epsilon)} = \frac{\epsilon}{1-\epsilon}$$

with probability at least $1 - 6\exp\left(-\frac{d\epsilon^2}{\alpha \log d}\right) - 2d^{-\alpha+1}$. Therefore, for $\alpha > 1$, the probability of having $|\cos\theta_d| \ge \frac{\epsilon}{1-\epsilon}$ goes to 0 as $d \to \infty$ for any $0 < \epsilon < 1$. So for $\boldsymbol{x}, \boldsymbol{y}$ drawn from Gaussian distribution, they also are likely to be orthogonal as $d \to \infty$.

Here is an even simpler approach for two Gaussian random vectors. Since the normal distribution is invariant under rotations a Gaussian random vector remains Gaussian if we apply a unitary matrix. Thus we can assume w.l.o.g. that $\boldsymbol{x}$ is the first unit vector (normalized such that it has norm $\sqrt{d}$) . Thus, $\langle \boldsymbol{x}, \boldsymbol{y}\rangle = y_1$. Since y_1 is just a scalar Gaussian random variable, the result follows now easily.

Problem 2 Consider the following setup. Given a square of side length 1, we place four circles in the square as depicted in Figure 1 (each of the gray circles has radius 1/4). We now place a circle at the center of the square (the blue circle in Figure 1) such that this circle in the middle touches each of the four identical circles. Let r denote the radius of the blue circle.

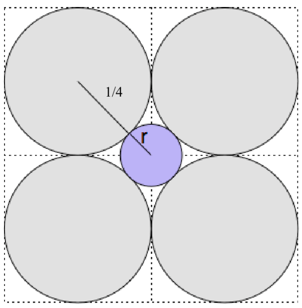

Figure 1: 4 circles

We can do something analogous in three dimensions, see Figure 2. We place eight spheres of radius 1/4 inside a cube of side length 1, and put a (blue) sphere in the middle such that it touches all eight (gray) spheres.

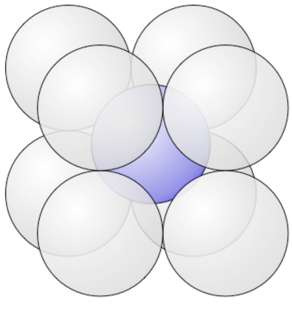

Figure 2: 8 spheres

In four dimensions we can arrange 16 hyperspheres of radius 1/4 inside a hypercube of side length 1 and place a hypersphere in the middle, so that this hypersphere all the other 16 hyperspheres.

Obviously we can do this for increasing dimension d. What happens with the blue hypersphere in the middle as d increases? Will it shrink? Will it be of constant size? Will it grow outside the hypercube?

(Hint: Check the diameter of the blue hypersphere in comparison to the sidelength of the cube as d increases. This is actually not difficult to compute, it may sound more complicated than it is).

Answer: An unit cube in $\mathbb{R}^d$ is $[-1/2, 1/2]^d$. In dimension $d \geq 2$, there are totally 2^d hyperspheres of radius 1/4. The centers of those hyperspheres are $\{-1/4, 1/4\}^d$. The ball centered at origin and touching all the spheres yields

$$r + \frac{1}{4} = \frac{\sqrt{d}}{4} \Longrightarrow r = \frac{\sqrt{d} - 1}{4}$$

since $\{-1/4, 1/4\}^d$, the origin and the touching point should be on one line.

In other words, if $r = \frac{1}{2}$, i.e., $n = 9$, the ball will touch the boundary of unit cube, and for $n > 9$ the ball will break through the faces of the unit cube.

Problem 3 Show that for every fixed dimension reduction matrix A of size $k \times d$ with $k < d$, there exists vectors $x, y \in \mathbb{R}^d$ such that the distance $\|Ax - Ay\|$ (no matter which norm we use) is vastly different from $\|x - y\|$.

Answer: Remember that the dimension reduction matrix $A \in \mathbb{R}^{k \times d}$ with $k < d$ has a nontrivial null space, i.e., there exists a $z \in \mathbb{R}^d$ such that $Az = 0 \in \mathbb{R}^k$ with $z \neq 0$. Without loss of generality, we just assume $\|z\| = 1$. Now we let x and y satisfy $x - y = c_0 z$ for some large constant c_0. Obviously, $\|A(x - y)\| = 0$ and on the other hand, $\|c_0 z\| = c_0$ is large.

Problem 4 The Yale Face Database contains images from various individuals in different poses and under different lighting conditions. Some of the images are stored in the file `SomeYaleFaces.mat`.

Load this file into Matlab. The variable `X` is a matrix of size 1024×2414. Each column of `X` is an image of size 32×32 (in vectorized form). The 2414 columns are images of 38 different persons in about 64 poses each. You can easily covert the k-th column of `X` back to an image via the commands

```
xk = X(:,k); xk = reshape(xk,32,32);
```

The command

```
imagesc(x1); colormap(gray);
```

will display the image.

You can conveniently display multiple images if you want with the file `showfaces.m`.

We want to compare three dimension reduction methods by comparing how well distances between the different images are preserved: Johnson-Lindenstrauss projection, Fast Johnson Lindenstrauss projection and simple random sampling (i.e., randomly picking k indices).

Choose different values for the reduced dimension k and compare the dimension reduction ability of the three methods. You need to think about how to devise such an experiment. There are of course multiple options to do so.

Answer: Here we use three methods to do dimension reduction. We measure the performance via

$$\epsilon_k = \max_{1\leq i\neq j\leq n}\left|\frac{\|f(\boldsymbol{x}_i)-f(\boldsymbol{x}_j)\|^2}{\|\boldsymbol{x}_i-\boldsymbol{x}_j\|^2}-1\right|$$

where $f(\boldsymbol{x}_i)\in\mathbb{R}^k$. That's the largest pairwise relative error after mapping $\boldsymbol{x}_i$ onto $\mathbb{R}^k$. Figure 3 show how $\log(\epsilon_k)$ depends on k for three different methods. One can easily see that the performances of random projection and fast JL transforms are quite similar in terms of relative error. However, random sampling does a poor job.

Note that theoretically, $k=\mathcal{O}(\epsilon_k^{-2})$ for both random projection and fast JL, so it is supposed that $k\epsilon_k^2$ should scale with $\mathcal{O}(1)$ if d and n are fixed. This seems valid in Figure 4 where $k\epsilon_k^2$ is quite small (around $10\approx\log(n)=\log(2414)$) and does not change much w.r.t. k. On the other hand, $k\epsilon_k^2$ fluctuates quite a lot for random sampling.

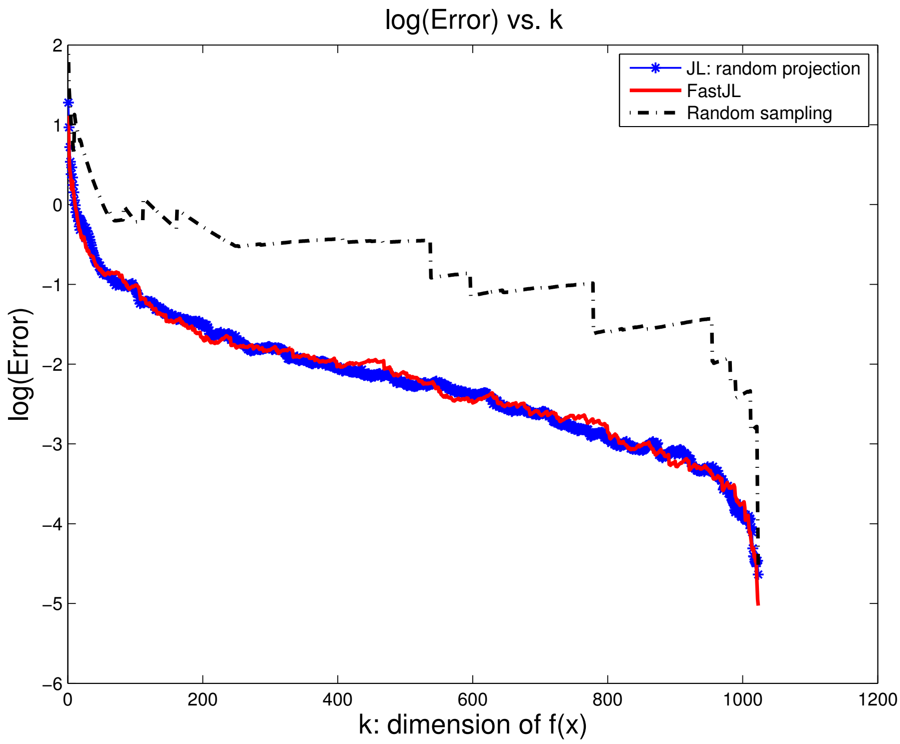

Figure 3: $\log(\epsilon_k)$ vs. k

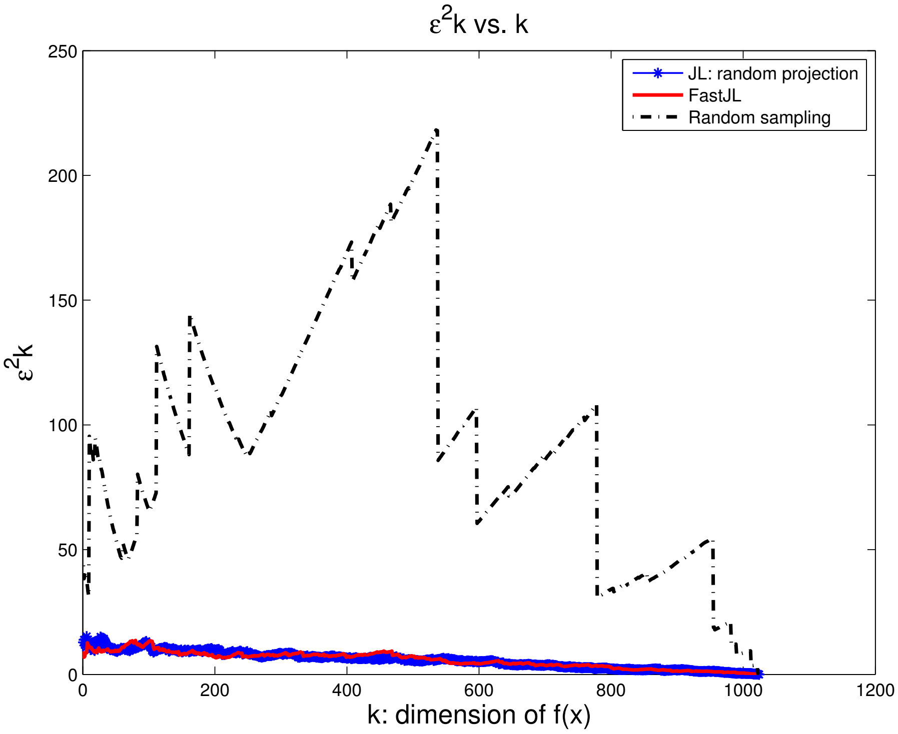

Figure 4: $k\epsilon_k^2$ vs. k