IE531: Algorithms for Data Analytics
Spring, 2022
Homework 4: Random Projections
Due Date: March 24, 2023

Let X be a discrete random variable that can take on three values in the set $\{-\sqrt{3}, 0, \sqrt{3}\}$, where

$$Prob\{X = -\sqrt{3}\} = Prob\{X = \sqrt{3}\} = \frac{1}{6}, \text{ and } Prob\{X = 0\} = \frac{2}{3}.$$

Let $\{\mathbf{u}_i\}_{i=1}^k$ be a set of k-many, $(d \times 1)$-vectors that resulted from making d-many calls to code that generates i.i.d. samples of the RV X. Assume d is large, and $k \ll d$.

Before we proceed, it should not be difficult for you to see that

$$\begin{aligned}
E\{X\} &= \cancel{\left(-\sqrt{3} \times \frac{1}{6}\right)} + \left(0 \times \frac{2}{3}\right) + \cancel{\left(\sqrt{3} \times \frac{1}{6}\right)} = 0 \\
var\{X\} &= E\{(X - \underbrace{E\{X\}}_{=0})^2\} = E\{X^2\} = \left((-\sqrt{3})^2 \times \frac{1}{6}\right) + \left(0^2 \times \frac{2}{3}\right) + \left(\sqrt{3}^2 \times \frac{1}{6}\right) = 1
\end{aligned}$$

If $\mathbf{u}$ is a $(d \times 1)$-vector, whose elements are d-many i.i.d. samples of the RV X described above, then each component of $\mathbf{u}$ (i.e. $\mathbf{u}_i, i \in \{1, 2, \ldots, d\}$) can take on only one of three values – $\{-\sqrt{3}, 0, \sqrt{3}\}$. This means, we can look at the (random) vector $\mathbf{u}$ as

$$\begin{aligned}
\mathbf{u} &= \sqrt{3} \times \begin{pmatrix}
\left\{\underbrace{-1}_{\text{Prob. } \frac{1}{6}}, \underbrace{0}_{\text{Prob. } \frac{2}{3}}, \underbrace{1}_{\text{Prob. } \frac{1}{6}}\right\} \\
\left\{\underbrace{-1}_{\text{Prob. } \frac{1}{6}}, \underbrace{0}_{\text{Prob. } \frac{2}{3}}, \underbrace{1}_{\text{Prob. } \frac{1}{6}}\right\} \\
\vdots \\
\left\{\underbrace{-1}_{\text{Prob. } \frac{1}{6}}, \underbrace{0}_{\text{Prob. } \frac{2}{3}}, \underbrace{1}_{\text{Prob. } \frac{1}{6}}\right\}
\end{pmatrix} \\
\frac{\|\mathbf{u}\|_2^2}{3} &= \sum_{i=1}^{d} \left\{\underbrace{(-1)^2}_{\text{Prob. } \frac{1}{6}}, \underbrace{0^2}_{\text{Prob. } \frac{2}{3}}, \underbrace{1^2}_{\text{Prob. } \frac{1}{6}}\right\} \\
&= \sum_{i=1}^{d} \left\{\underbrace{0}_{\text{Prob. } \frac{2}{3}}, \underbrace{1}_{\text{Prob. } \frac{1}{3}}\right\}
\end{aligned}$$

That is, $\frac{\|\mathbf{u}\|_2^2}{3} \in \{0, 1, \ldots, d\}$ has a Binomial Distribution, where

$$Prob\left\{\frac{\|\mathbf{u}\|_2^2}{3} = i, i \in \{0, 1, \ldots, d\}\right\} = \left(\begin{array}{c} d \\ i \end{array}\right) \times \left(\frac{1}{3}\right)^i \times \left(\frac{2}{3}\right)^{d-i}$$

From your course in Probability and Statistics, you know the mean of this Binomial Distribution is $\frac{1}{3} \times d$, which means $E\{\|\mathbf{u}\|_2^2\} = \cancel{3} \times \frac{d}{\cancel{3}} = d \Rightarrow E\{\|\mathbf{u}\|_2\} = \sqrt{d}$. This can be established in other ways (as you will see below). I will let you work out what $var(\|\mathbf{u}\|_2^2)$ should be, from what you know about the Binomial Distribution. You could use this to arrive at a tail-bound that uses Chebyshev's Inequality. That said, you should also know that there are expressions for the tail-distribution of the Binomial Distribution that you can use in lieu of Chernoff Bounds, if you wish. In one sentence – we can "escape" from the high-theory of Spherical Gaussians if we are to use the distribution introduced to you in this homework.

The material presented above is about the nature of the k-many, $(d \times 1)$-vectors in the set $\{\mathbf{u}_i\}_{i=1}^k$. When we compute $\mathbf{y} = \frac{1}{\sqrt{k}}\mathbf{A}^T\mathbf{x}$, each row/component of the $\mathbf{y}$ vector is a linear-combination of d-many i.i.d. RVs from the distribution described above. That is, $\mathbf{y}_j = \frac{1}{\sqrt{k}} \times \sum_{i=1}^d \mathbf{x}_i \times c_i$, where $j \in \{1, 2, \ldots, k\}$, and c_i is a RV that takes on values in the set $\{\sqrt{3}, 0, \sqrt{3}\}$ with probabilities $(\frac{1}{6}, \frac{2}{3}, \frac{1}{6})$, respectively. It follows that

$$\begin{aligned}
\mathbf{y}_j^2 &= \frac{1}{k} \times \left(\sum_{i=1}^d \mathbf{x}_i^2 \times c_i^2\right) + \left(\sum_{i,k=1, i\neq k}^d \mathbf{x}_i \times \mathbf{x}_k \times c_i \times c_k\right) \\
\Rightarrow E\{\mathbf{y}_j^2\} &= \frac{1}{k} \times \left(\sum_{i=1}^d \mathbf{x}_i^2 \times \underbrace{E\{c_i^2\}}_{=1}\right) + \left(\sum_{i,k=1, i\neq k}^d \mathbf{x}_i \times \mathbf{x}_k \times \underbrace{E\{c_i \times c_k\}}_{=0}\right) \\
\Rightarrow E\{\mathbf{y}_j^2\} &= \frac{1}{k} \times \sum_{i=1}^d \mathbf{x}_i^2 = \frac{1}{k} \times \|\mathbf{x}\|_2^2
\end{aligned}$$

The expected/average value of the length-squared of $\mathbf{y}$ is the expected/average value of the sum each of its k-many components. That is,

$$E\{\|\mathbf{y}\|_2^2\} = \sum_{j=1}^k E\{\mathbf{y}_j^2\} = \sum_{j=1}^k \left(\frac{1}{k} \times \|\mathbf{x}\|_2^2\right) = \|\mathbf{x}\|_2^2.$$

We can do even better, but I will refrain from getting into the nitty-gritty here. There are only a finite-set of discrete-values that $\mathbf{y}_j^2$ (see above derivation) can take. To illustrate this, let us look at the simple case of $d = 2$, that is $\mathbf{y}_j = (\mathbf{x}_1 \times c_1 + \mathbf{x}_2 \times c_2)$, this would mean $\mathbf{y}_j^2 = (\mathbf{x}_1^2 c_1^2 + \mathbf{x}_2^2 c_2^2 + 2\mathbf{x}_1\mathbf{x}_2 c_1 c_2)$, where $c_1, c_2 \in \{-\sqrt{3}, 0, \sqrt{3}\}$. You can try all nine combinations, and you will get the result that $\mathbf{y}_j^2 \in \{3\mathbf{x}_1^2 + 3\mathbf{x}_2^2 + 6\mathbf{x}_1\mathbf{x}_2, 3\mathbf{x}_1^2 + 3\mathbf{x}_2^2 - 6\mathbf{x}_1\mathbf{x}_2, 3\mathbf{x}_1^2, 3\mathbf{x}_2^2, 0\}$, and it takes on these values with probability $(\frac{1}{18}, \frac{1}{18}, \frac{4}{18}, \frac{4}{18}, \frac{8}{18})$, respectively. You can verify that

$$E\{\mathbf{y}_j^2\} = \left((3\mathbf{x}_1^2 + 3\mathbf{x}_2^2 + 6\mathbf{x}_1\mathbf{x}_2) \times \frac{1}{18}\right) + \left((3\mathbf{x}_1^2 + 3\mathbf{x}_2^2 - 6\mathbf{x}_1\mathbf{x}_2) \times \frac{1}{18}\right) + \left(3\mathbf{x}_1^2 \times \frac{4}{18}\right) + \left(3\mathbf{x}_2^2 \times \frac{4}{18}\right) = \mathbf{x}_1^2 + \mathbf{x}_2^2 = \|\mathbf{x}\|_2^2$$

Repeat this for the general case where d can take on any value – you will get these results (a) there are only a finite number of possible values for $\mathbf{y}_j^2$, and (b) you can compute the probabilities that $\mathbf{y}_j^2$ takes on each of these finite-number of values with relative ease. This would mean you can get compute the tail-probabilities very accurately for this case. This is in sharp contrast to the case when we use Spherical Gaussian RVs used in the JL-Lemma – where $\mathbf{y}_j^2$ can take on an infinite-set of possible values (with different probabilities, off course; the chance that we encounter a super-large value is very very rare, etc; this is the basis of the Gaussian Annulus Theorem etc).

1. (*25 points*) Show that $||\mathbf{u}_1||_2 = \sqrt{d}$ with high-probability.

   See above – $E\{||\mathbf{u}||_2\} = \sqrt{d}$. You could also just use the Law of Large Numbers (i.e. the proof in the Book & my Jupyter Notebook) to show that

$$Prob\left(\left|\left(\frac{1}{d}\sum_{i=1}^{d}\mathbf{u}_i^2\right) - \overbrace{E\{\mathbf{u}_i^2\}}^{=1}\right| \geq \epsilon\right) \leq \frac{\overbrace{Var\{\mathbf{u}_i^2\}}^{=1}}{4d\epsilon^2},$$

   This means that for large d, $||\mathbf{u}||_2^2 \approx d$.

2. (*25 points*) Show that the members of the set $\{\mathbf{u}_i\}_{i=1}^k$ are mutually-perpendicular with high-probability.

   Verbatim repeat of the material in the Book/Lesson, with the observations above, should do the needful here. Just for the record – if $\mathbf{u}_i, \mathbf{u}_j$ are two vectors from the set described above. We have

$$||\mathbf{u}_i - \mathbf{u}_j||_2^2 = \sum_{k=1}^{d}(\mathbf{u}_{i_k} - \mathbf{u}_{j_k})^2$$

   Since these vectors have been constructed by independent calls to the RV-generator for X, it follows that $E\{(\mathbf{u}_{i_k} - \mathbf{u}_{j_k})^2\} = 2$. We conclude that (for large d) $\sum_{k=1}^{d}(\mathbf{u}_{i_k} - \mathbf{u}_{j_k})^2 = 2d$, almost surely. From $\mathbf{c} = \mathbf{u}_i - \mathbf{u}_j$, it follows from the *Law of Cosines*[1] that if θ is the angle between $\mathbf{u}_i$ and $\mathbf{u}_j$, then

$$\underbrace{||\mathbf{c}||^2}_{=2d} = \underbrace{||\mathbf{u}_i||^2}_{=d} + \underbrace{||\mathbf{u}_j||^2}_{=d} - \underbrace{(2||\mathbf{u}_i|| \times ||\mathbf{u}_j||Cos(\theta))}_{\Rightarrow Cos(\theta)=0}.$$

   This would mean that $\mathbf{u}_i$ and $\mathbf{u}_j$ are orthogonal to each other almost surely (for large d).

3. (*25 points*) Let $\mathbf{x}$ be a $(d \times 1)$-vector and $\mathbf{y}$ be a $(k \times 1)$-vector, where

$$\mathbf{y} = \frac{1}{\sqrt{k}}\mathbf{A}^T\mathbf{x}, \text{ where } \mathbf{A} = \begin{pmatrix} \mathbf{u}_1 & \mathbf{u}_2 & \cdots & \mathbf{u}_k \end{pmatrix}$$

   Show that $E\{||\mathbf{y}||_2^2\} = ||\mathbf{x}||_2^2$.

[1] See section 3 of the Review of Linear Algebra.

$$\begin{aligned}
\|\mathbf{y}\|_2^2 &= \mathbf{y}^T\mathbf{y} = \frac{1}{k}\mathbf{x}^T\mathbf{A}\mathbf{A}^T\mathbf{x} = \frac{1}{k}\mathbf{x}^T \begin{pmatrix} \mathbf{u}_1 & \mathbf{u}_2 & \cdots & \mathbf{u}_k \end{pmatrix} \begin{pmatrix} \mathbf{u}_1^T \\ \mathbf{u}_2^T \\ \cdots \\ \mathbf{u}_k^T \end{pmatrix} \mathbf{x} \\
&= \frac{1}{k}\mathbf{x}^T \left(\sum_{i=1}^{k} \mathbf{u}_i \times \mathbf{u}_i^T \right) \mathbf{x} \\
\Rightarrow E\{\|\mathbf{y}\|_2^2\} &= \frac{1}{k}\mathbf{x}^T \underbrace{\left(\sum_{i=1}^{k} \underbrace{E\{\mathbf{u}_i \times \mathbf{u}_i^T\}}_{I_{k\times k}} \right)}_{k \times I_{k\times k}} \mathbf{x} = \frac{1}{\cancel{k}}\mathbf{x}^T \left(\cancel{k} \times I_{k\times k} \right) \mathbf{x} = \mathbf{x}^T\mathbf{x} = \|\mathbf{x}\|_2^2
\end{aligned}$$

4. (*25 points*) Assume there is a constant C such that

$$Prob\{\left| \, \|\mathbf{y}\|_2^2 - \|\mathbf{x}\|_2^2 \, \right| > \epsilon\} < e^{-Ck\epsilon^2}$$

Show that any value of

$$k > \frac{1}{C\epsilon^2} \ln \frac{1}{(1-\delta)}$$

is sufficient to guarantee that if two d-dimensional vectors are close to each other, then their k-dimensional "proxies" will be close to each other with probability δ.

$$e^{-Ck\epsilon^2} < (1-\delta) \Rightarrow -Ck\epsilon^2 < \ln(1-\delta) \Rightarrow Ck\epsilon^2 > -\ln(1-\delta)(= \ln(\frac{1}{1-\delta})) \Rightarrow k > \frac{1}{C\epsilon^2} \ln \frac{1}{(1-\delta)}$$

If $\mathbf{x}_1$ and $\mathbf{x}_2$ are two $(d \times 1)$ vectors that are close to each other, we expect the Euclidean Norm of the $(d \times 1)$-vector $\mathbf{a} = \mathbf{x}_1 - \mathbf{x}_2$ to be small. The $(k \times 1)$-proxy of $\mathbf{a}$ is essentially $\mathbf{b} = \mathbf{y}_1 - \mathbf{y}_2$, where $\mathbf{y}_1$ (resp. $\mathbf{y}_2$) is the $(k \times 1)$-proxy of $\mathbf{x}_1$ (resp. $\mathbf{x}_2$). For the above value of k we know that $\|\mathbf{b}\|_2^2 \approx \|\mathbf{a}\|_2^2 \approx 0$. For the chosen k,

$$Prob\{\left| \, \|\mathbf{a}\|_2^2 - \|\mathbf{b}\|_2^2 \, \right| > \epsilon\} < 1 - \delta \Rightarrow Prob\{\left| \, \|\mathbf{a}\|_2^2 - \|\mathbf{b}\|_2^2 \, \right| < \epsilon\} > \delta$$