Machine Learning for Data Science (CS 4786)

Lecture 5: Random Projections

The text in black outlines high level ideas. The text in blue provides simple mathematical details to "derive" or get to the algorithm or method. The text in red are mathematical details for those who are interested.

1 Motivation

Consider the scenario when both dimensionality d of the original dataset and the number of data points n are large. Specifically in this scenario, PCA is not a viable option as it is computationally intensive and in fact requires to store/access the entire data matrix which can be prohibitive. Random projection is catered to exactly such scenario and works by producing matrix W, the $d \times K$ projection matrix by randomly sampling its entries!

2 Random Projection Method

The random projection method constitutes of sampling the projection matrix W by filling each of its entries by sampling independently at random as:

$$W[i,j] = \begin{cases} +\frac{1}{\sqrt{K}} & \text{with probability } 1/2 \\ -\frac{1}{\sqrt{K}} & \text{with probability } 1/2 \end{cases}$$

That is we flip a fair coin to set the sign of each entry to $+$ or $-$ of $1/\sqrt{K}$. The surprising fact is that random projection provides a guarantee that after projecting the points from the d dimensional space to K dimensional space using the randomly drawn matrix W, the distances between the high dimensional points is preserved in the lower dimensional projections up to some approximation factor.

Specifically, the following lemma known as the Johnson Lindenstrauss Lemma formalizes the guarantee of the random projection method.

Lemma 1. *For any $\delta > 0$ and $\epsilon > 0$, if we pick $K > \frac{20 \log(n/\delta)}{\epsilon^2}$ then, with probability at least $1 - \delta$ over the draw of random projection matrix W, for all $t, s \in \{1, \ldots, n\}$,*

$$(1-\epsilon)\|\mathbf{y}_t - \mathbf{y}_s\|^2 \leq \|\mathbf{x}_t - \mathbf{x}_s\|^2 \leq (1+\epsilon)\|\mathbf{y}_t - \mathbf{y}_s\|^2$$

where each $\mathbf{y}_t = \mathbf{x}_t^\top W$ is the K dimensional projection of corresponding point $\mathbf{x}_t \in \mathbb{R}^d$.

The above lemma is rather surprising at first glance, as the dimensionality d of the original data does not come into play in choice of K. Hence the same K holds for same choice of n, ϵ and δ, irrespective of whether the original dimensionality is 10 or a million or a 100-billion.

3 Why Random Projection Works?!

In this section, first we will provide at a high level, an explanation of why random projection could work by showing that random projection preserves distances in expectation. In subsequent subsection, we will provide the formal high probability version of the JL Lemma for interested readers.

3.1 Simpler In-Expectation Version

Consider any vector $\tilde{\mathbf{x}} \in \mathbb{R}^d$ and let $\tilde{\mathbf{y}} = \tilde{\mathbf{x}}^\top W$ be the random projection of $\tilde{\mathbf{x}}$. Note that:

$$\begin{aligned}
|\tilde{\mathbf{y}}[j]|^2 &= \left(\sum_{i=1}^{d} W[i,j] \cdot \tilde{\mathbf{x}}[i]\right)^2 \\
&= \sum_{i=1}^{d} \left(W[i,j] \cdot \tilde{\mathbf{x}}[i]\right)^2 + 2\sum_{i'>i} \left(W[i,j] \cdot \tilde{\mathbf{x}}[i]\right)\left(W[i',j] \cdot \tilde{\mathbf{x}}[i']\right) \\
&= \sum_{i=1}^{d} W^2[i,j]\tilde{\mathbf{x}}^2[i] + \sum_{i'>i} \left(W[i,j] \cdot W[i',j]\right) \cdot \left(\tilde{\mathbf{x}}[i] \cdot \tilde{\mathbf{x}}[i']\right)
\end{aligned}$$

However $W^2[i,j] = 1/K$ and so

$$= \frac{1}{K}\sum_{i=1}^{d} \tilde{\mathbf{x}}^2[i] + \sum_{i'>i} \left(W[i,j] \cdot W[i',j]\right) \cdot \left(\tilde{\mathbf{x}}[i] \cdot \tilde{\mathbf{x}}[i']\right)$$

Now note that $\mathbb{E}\left[(W[i,j] \cdot W[i',j])\right] = \mathbb{E}\left[W[i,j]\right] \cdot \mathbb{E}\left[W[i',j]\right] = 0$ and so,

$$\mathbb{E}\left[|\tilde{\mathbf{y}}[j]|^2\right] = \frac{1}{K}\sum_{i=1}^{d} \tilde{\mathbf{x}}^2[i] = \frac{1}{K}\|\tilde{\mathbf{x}}\|^2$$

Further, note that each term $(\tilde{\mathbf{x}}[i] \cdot \tilde{\mathbf{x}}[i'])$ is 0 mean and symmetric. Hence in fact, $|\tilde{\mathbf{y}}[j]|^2$ is symmetrically distributed with expected value of $\frac{1}{K}\|\tilde{\mathbf{x}}\|^2$.

From this we conclude that,

$$\mathbb{E}\left[\|\tilde{\mathbf{y}}\|^2\right] = \sum_{j=1}^{K} \mathbb{E}\left[|\tilde{\mathbf{y}}[j]|^2\right] = \sum_{j=1}^{K} \frac{1}{K}\|\tilde{\mathbf{x}}\|^2 = \|\tilde{\mathbf{x}}\|^2$$

Thus we see that the expected norm squared of $\tilde{\mathbf{y}}$ is same as norm squared of $\tilde{\mathbf{x}}$. Now further note that each $\tilde{\mathbf{y}}[j]$ is independent of $\tilde{\mathbf{y}}[j']$ for any $j, j' \in \{1, \ldots, K\}$. Thus we can think of $\|\tilde{\mathbf{y}}\|^2$ as an average of K independent random variables whose expectations are $\|\tilde{\mathbf{x}}\|^2$. Thus we can expect a high probability statement.

Now if we let $\tilde{\mathbf{x}} = \mathbf{x}_t - \mathbf{x}_s$, then

$$\tilde{\mathbf{y}} = (\mathbf{x}_t - \mathbf{x}_s)^\top W = \mathbf{x}_t^\top W - \mathbf{x}_s^\top W = \mathbf{y}_t - \mathbf{y}_s \ .$$

Thus the expected distance square between $\mathbf{y}_t$ and $\mathbf{y}_s$ is same as distance between $\mathbf{x}_t$ and $\mathbf{x}_s$. In the following section we shall in fact prove the high probability JL lemma.

3.2 High Probability Version

To prove the high probability version of this statement, we need the following simple lemma.

Lemma 2 (Hoeffding Lemma (paraphrased)). *Let X be a random variable that takes value c_i with probability $1/2$ and value $-c_i$ with probability $1/2$. Then, we have that for any λ*

$$\mathbb{E}\left[\exp\left(\lambda X\right)\right] \leq \exp\left(\lambda^2 c^2/2\right)$$

Proof.

$$\mathbb{E}\left[\exp\left(\lambda X\right)\right] = \frac{1}{2}\left(\exp\left(\lambda c_i\right) + \exp\left(-\lambda c_i\right)\right) \leq \exp\left(\lambda^2 c_i^2/2\right)$$

where we use the fact that for any a, $e^a + e^{-a} \leq 2e^{a^2/2}$ □

Now we use the above lemma to obtain the following proposition.

Lemma 3. *Consider any vector $\tilde{\mathbf{x}} \in \mathbb{R}^d$ and let $\tilde{\mathbf{y}} = \tilde{\mathbf{x}}^\top W$ be the random projection of $\tilde{\mathbf{x}}$. We have that for any $j \in \{1, \ldots, K\}$, and any λ,*

$$\mathbb{E}\left[\exp\left(\lambda \tilde{\mathbf{y}}[j]\right)\right] \leq \exp\left(\frac{\lambda^2 \|\tilde{\mathbf{x}}\|^2}{2K}\right)$$

and further, for any $s \leq K/\|\tilde{\mathbf{x}}\|^2$

$$\mathbb{E}\left[\exp\left(s\|\tilde{\mathbf{y}}\|^2/2\right)\right] \leq \left(1 - \frac{s\|\tilde{\mathbf{x}}\|^2}{K}\right)^{-1/2}$$

Proof. Consider any vector $\tilde{\mathbf{x}} \in \mathbb{R}^d$ and let $\tilde{\mathbf{y}} = \tilde{\mathbf{x}}^\top W$ be the random projection of $\tilde{\mathbf{x}}$. Now, consider any $j \in \{1, \ldots, K\}$, the term $\tilde{\mathbf{y}}[j] = \sum_{i=1}^d \tilde{\mathbf{x}}[i]W[i,j]$ can be considered as sum of d independently distributed random variables $\tilde{\mathbf{x}}[1]W[1,j], \ldots, \tilde{\mathbf{x}}[d]W[d,j]$. Further, each $\tilde{\mathbf{x}}[i]W[i,j]$ is $\pm\tilde{\mathbf{x}}[i]|/\sqrt{K}$ with equal probability. Hence, using Hoeffding's inequality, for any λ,

$$\begin{aligned}
\mathbb{E}\left[\exp\left(\lambda \tilde{\mathbf{y}}[j]\right)\right] &= \mathbb{E}\left[\exp\left(\lambda \sum_{i=1}^d \tilde{\mathbf{x}}[i]W[i,j]\right)\right] \\
&= \prod_{i=1}^d \mathbb{E}\left[\exp\left(\lambda \tilde{\mathbf{x}}[i]W[i,j]\right)\right]
\end{aligned}$$

Using Hoeffding Lemma, with $X_i = \tilde{\mathbf{x}}[i]W[i,j]$

$$\begin{aligned}
&\leq \prod_{i=1}^d \exp\left(\lambda^2 \left(\frac{\mathbf{x}[i]}{\sqrt{K}}\right)^2/2\right) \\
&= \exp\left(\lambda^2 \sum_{i=1}^d \left(\frac{\mathbf{x}[i]}{\sqrt{K}}\right)^2/2\right) \\
&= \exp\left(\frac{\lambda^2 \|\tilde{\mathbf{x}}\|^2}{2K}\right)
\end{aligned}$$

Hence we conclude that for any λ,

$$\mathbb{E}\left[\exp\left(\lambda\tilde{\mathbf{y}}[j]\right)\right] \leq \exp\left(\frac{\lambda^2\|\tilde{\mathbf{x}}\|^2}{2K}\right) \tag{1}$$

This yields the first inequality. To prove the second inequality, we use a really cool trick with Gaussian random variables. Consider the random variable $Z \sim N(0,1)$, a standard normal random variable. For such a variable, using the moment generating function, we have that for any α, $\mathbb{E}\left[\exp(\alpha Z)\right] = \exp(\alpha^2/2)$. Specifically using this with $\alpha = \sqrt{s}\tilde{\mathbf{y}}[j]$, we can conclude that $\exp\left(s\tilde{\mathbf{y}}[j]^2/2\right) = \mathbb{E}_Z\left[\exp\left(\sqrt{s}\tilde{\mathbf{y}}[j]Z\right)\right]$. Hence, for any $s > 0$

$$\begin{aligned}\mathbb{E}\left[\exp\left(s\tilde{\mathbf{y}}[j]^2/2\right)\right] &= \mathbb{E}\left[\mathbb{E}_Z\left[\exp\left(\sqrt{s}\tilde{\mathbf{y}}[j]Z\right)\right]\right] \\ &= \mathbb{E}_Z\left[\mathbb{E}\left[\exp\left((\sqrt{s}Z)\tilde{\mathbf{y}}[j]\right)\right]\right]\end{aligned}$$

Using inequality in Eq. 1 with $\lambda = (\sqrt{s}Z)$ we get,

$$\begin{aligned}&\leq \mathbb{E}_Z\left[\exp\left(\frac{sZ^2\|\tilde{\mathbf{x}}\|^2}{2K}\right)\right] \\ &= \left(1 - \frac{s\|\tilde{\mathbf{x}}\|^2}{K}\right)^{-1/2}\end{aligned}$$

The last equality holds for any s s.t. $\frac{s\|\tilde{\mathbf{x}}\|^2}{2K} \leq 1/2$ and is obtained because Z^2 follows a Chi-square distribution and we are using the moment generating function for this distribution. Hence we have that $K \geq s\|\tilde{\mathbf{x}}\|^2$,

$$\mathbb{E}\left[\exp\left(s\tilde{\mathbf{y}}[j]^2/2\right)\right] \leq \left(1 - \frac{s\|\tilde{\mathbf{x}}\|^2}{K}\right)^{-1/2}$$

Now $\tilde{\mathbf{y}}[j]$'s are independently drawn and so,

$$\mathbb{E}\left[\exp\left(s\|\tilde{\mathbf{y}}\|^2/2\right)\right] = \mathbb{E}\left[\exp\left(s\sum_{j=1}^{K}\tilde{\mathbf{y}}[j]^2/2\right)\right] = \prod_{j=1}^{K}\mathbb{E}\left[\exp\left(s\tilde{\mathbf{y}}[j]^2/2\right)\right] \leq \left(1 - \frac{s\|\tilde{\mathbf{x}}\|^2}{K}\right)^{-K/2}$$

This conclude the proof. □

Now we are in a position to prove the JL lemma.

Lemma 4 (JL Lemma)**.** *For any $\delta > 0$ and $\epsilon > 0$, if we pick $K > \frac{12\log(n(n-1))}{\epsilon^2}$, with probability $1-\delta$, for any $s, t \in \{1, \ldots, n\}$*

$$(1-\epsilon)\|\mathbf{x}_t - \mathbf{x}_s\|^2 \leq \|\mathbf{y}_t - \mathbf{y}_s\|^2 \leq (1+\epsilon)\|\mathbf{x}_t - \mathbf{x}_s\|^2$$

Proof.

$$P\left(\|\tilde{\mathbf{y}}\|^2 > (1+\epsilon)\|\tilde{\mathbf{x}}\|^2\right) = P\left(\exp(s\|\tilde{\mathbf{y}}\|^2/2) > \exp(s(1+\epsilon)\|\tilde{\mathbf{x}}\|^2/2)\right)$$

However by Markov inequality, $P(X > \theta) \leq E[X]/\theta$ for any non-negative R.V. X and so

$$\leq \frac{\mathbb{E}\left[\exp(s\|\tilde{\mathbf{y}}\|^2/2)\right]}{\exp(s(1+\epsilon)\|\tilde{\mathbf{x}}\|^2/2)}$$

using previous lemma for the numerator,

$$\leq \left(1 - \frac{s\|\tilde{\mathbf{x}}\|^2}{K}\right)^{-K/2} \exp\left(-s(1+\epsilon)\|\tilde{\mathbf{x}}\|^2/2\right)$$

Using $s = \epsilon K/(1+\epsilon)\|\tilde{\mathbf{x}}\|^2$ which satisfies the condition that $s \leq K/\|\tilde{\mathbf{x}}\|^2$, we get,

$$\begin{aligned} P\left(\|\tilde{\mathbf{y}}\|^2 > (1+\epsilon)\|\tilde{\mathbf{x}}\|^2\right) &\leq \left(1 - \frac{\epsilon}{1+\epsilon}\right)^{-K/2} \exp\left(-\epsilon K/2\right) \\ &= (1+\epsilon)^{K/2} \exp\left(-\epsilon K/2\right) \end{aligned}$$

using the fact that $1 + x \leq \exp\left(x - x^2/2 + x^3/3\right)$ we conclude that,

$$\leq \exp\left(-\epsilon^2 K/4 + \epsilon^3 K/6\right)$$

Since $\epsilon < 1$, we have that $-\epsilon^2 K/4 + \epsilon^3 K/6 < -\epsilon^2 K/12$ and hence,

$$\leq \exp\left(-\epsilon^2 K/12\right)$$

Now we use the fact that $\tilde{\mathbf{y}}^2[j]$ is symmetrically distributed about its mean and so

$$P\left(\|\tilde{\mathbf{y}}\| > (1+\epsilon)\|\tilde{\mathbf{x}}\| \text{ or } \|\tilde{\mathbf{y}}\| < (1-\epsilon)\|\tilde{\mathbf{x}}\|\right) \leq 2\exp\left(-\epsilon^2 K/12\right)$$

What we have show so far is that, by setting $\tilde{\mathbf{x}} = \mathbf{x}_t - \mathbf{x}_s$, we have that, for any s, t,

$$P\left(\|\mathbf{y}_t - \mathbf{y}_s\| > (1+\epsilon)\|\mathbf{x}_t - \mathbf{x}_s\| \text{ or } \|\mathbf{y}_t - \mathbf{y}_s\| < (1-\epsilon)\|\mathbf{x}_t - \mathbf{x}_s\|\right) \leq 2\exp\left(-\epsilon^2 K/12\right)$$

Hence using union bound over all the $n(n-1)/2$ pairs $s, t,$, we have that,

$$P\left(\exists t, s \text{ s.t. } \|\mathbf{y}_t - \mathbf{y}_s\| > (1+\epsilon)\|\mathbf{x}_t - \mathbf{x}_s\| \text{ or } \|\mathbf{y}_t - \mathbf{y}_s\| < (1-\epsilon)\|\mathbf{x}_t - \mathbf{x}_s\|\right) \leq n(n-1)\exp\left(-\epsilon^2 K/12\right)$$

Hence if $K > \frac{12\log(n(n-1))}{\epsilon^2}$, we can conclude that

$$P\left(\exists t, s \text{ s.t. } \|\mathbf{y}_t - \mathbf{y}_s\| > (1+\epsilon)\|\mathbf{x}_t - \mathbf{x}_s\| \text{ or } \|\mathbf{y}_t - \mathbf{y}_s\| < (1-\epsilon)\|\mathbf{x}_t - \mathbf{x}_s\|\right) \leq \delta$$

The lemma statement is just another way of writing the above. □