The Random Projection Method *

Edo Liberty †

September 25, 2007

1 Introduction

We start by giving a short proof of the Johnson-Lindenstrauss lemma due to P.Indyk and R.Motowani. We continue by showing an application of random projections for the use of fast matrix rank k approximation due to Papadimitriou, Raghavan, Tamaki, and Vempala.

2 Random Projection

In the next 20 minutes we will prove the Johnson-Lindenstrauss (JL) lemma. We use a construction and proof derivation by Indyk and Motowani which are simpler then the original proof. Let us state the JL lemma:

Lemma 2.1 *A set of n points $\mathbf{u}_1 \dots \mathbf{u}_n$ in $\mathbb{R}^d$ can be projected down to $\mathbf{v}_1 \dots \mathbf{v}_n$ in $\mathbb{R}^k$ such that all pairwise distances are preserved:*

$$(1-\epsilon)\|\mathbf{u}_i - \mathbf{u}_j\|^2 \leq \|\mathbf{v}_i - \mathbf{v}_j\|^2 \leq (1+\epsilon)\|\mathbf{u}_i - \mathbf{u}_j\|^2$$

if

$$k > \frac{9 \ln n}{\epsilon^2 - \epsilon^3}, \textit{ and}, 0 \leq \epsilon \leq 1/2$$

In order to see this we first need to consider the probability that one vector $\mathbf{u}$ is distorted by more then ϵ. Our projecting matrix R will be a random $d \times k$ matrix s.t $R(i,j)$ are drawn *i.i.d* from $N(0,1)$[1]. We will see that if we set:

$$v = \frac{1}{\sqrt{k}} R^T \mathbf{u}$$

Then:

$$E(\|\mathbf{v}\|^2) = \|\mathbf{u}\|^2$$

And

$$(1-\epsilon)\|\mathbf{u}\|^2 \leq \|\mathbf{v}\|^2 \leq (1+\epsilon)\|\mathbf{u}\|^2$$

with probability $\Pr_{success} \geq 1 - 2e^{-(\epsilon^2 - \epsilon^3)\frac{k}{4}}$.

*chosen chapters from DIMACS vol.65 by Santosh S. Vempala

† Yale University, New Haven CT (edo.liberty@yale.edu) Supported by AFOSR, and NGA.

[1] Normal distribution with mean 0 and variance 1.

2.1 Calculating $E(\|\mathbf{v}\|^2)$

Since $\mathbf{v} = \frac{1}{\sqrt{k}} R^T \mathbf{u}$ we can calculate the expectancy $E(\|v\|^2)$.

$$
\begin{aligned}
E(\|\mathbf{v}\|^2) &= E\left(\sum_{i=1}^{k}\left(\sum_{j=1}^{d}\frac{1}{\sqrt{k}}R(i,j)\mathbf{u}(j)\right)^2\right) \\
&= \sum_{i=1}^{k}\frac{1}{k}E\left(\left(\sum_{j=1}^{d}R(i,j)\mathbf{u}(j)\right)^2\right) \\
&= \sum_{i=1}^{k}\frac{1}{k}\sum_{j=1}^{d}E\left(R(i,j)^2\right)\left(\mathbf{u}(j)^2\right) \\
&= \sum_{i=1}^{k}\frac{1}{k}\sum_{j=1}^{d}(\mathbf{u}(j))^2 \\
&= \|\mathbf{u}\|^2.
\end{aligned}
$$

2.2 Bounding the "stretching" probability

In order to bound the distortion probability we define $x_j = \frac{1}{\|\mathbf{u}\|}\langle R_j, \mathbf{u}\rangle$ and

$$
x = \frac{k\|\mathbf{v}\|^2}{\|\mathbf{u}\|^2} = \sum_{j=1}^{k}\frac{(R_j^T\mathbf{u})^2}{\|\mathbf{u}\|^2} = \sum_{j=1}^{k}x_j^2
$$

The convince in these definitions will become clear in the next few steps. We now turn to calculate the probability

$$
\begin{aligned}
\Pr\left[\|\mathbf{v}\|^2 \ge (1+\epsilon)\|\mathbf{u}\|^2\right] &= \Pr\left[x > (1+\epsilon)k\right] \\
&= \Pr\left[x \ge (1+\epsilon)k\right] \\
&= \Pr\left[e^{\lambda x} \ge e^{\lambda(1+\epsilon)k}\right] \\
&\le \frac{E(e^{\lambda x})}{e^{\lambda(1+\epsilon)k}} \\
&= \frac{\Pi_{j=1}^{k}E(e^{\lambda x_j^2})}{e^{\lambda(1+\epsilon)k}} \\
&= \left(\frac{E(e^{\lambda x_1^2})}{e^{\lambda(1+\epsilon)}}\right)^k
\end{aligned}
$$

Here we are faced with calculating $E(e^{\lambda x_1^2})$. Using the fact that x_1 itself is drawn from $N(0,1)$ We get that:

$$
\begin{aligned}
E(e^{\lambda x_1^2}) &= \int_{-\infty}^{\infty} e^{\lambda t^2}\frac{1}{\sqrt{2\pi}}e^{-\frac{t^2}{2}}dt \\
&= \frac{1}{\sqrt{1-2\lambda}}\int_{-\infty}^{\infty}\frac{\sqrt{1-2\lambda}}{\sqrt{2\pi}}e^{-\frac{t^2}{2}(1-2\lambda)}dt \\
&= \frac{1}{\sqrt{1-2\lambda}} \quad \text{For } \lambda < 1/2.
\end{aligned}
$$

Inserting $E(e^{\lambda X_1^2}) = \frac{1}{\sqrt{1-2\lambda}}$ into the probability equation we get:

$$
\Pr\left[\|\mathbf{v}\| \ge (1+\epsilon)\|\mathbf{u}\|^2\right] \le \left(\frac{e^{-2(1+\epsilon)\lambda}}{1-2\lambda}\right)^{k/2}
$$

Substituting for $\lambda = \frac{\epsilon}{2(1+\epsilon)}$ we get:

$$\Pr\left[\|\mathbf{v}\| \geq (1+\epsilon)\|\mathbf{u}\|^2\right] \leq \left((1+\epsilon)e^{-\epsilon}\right)^{k/2}$$

Finally using that $1+\epsilon < e^{\epsilon-(\epsilon^2-\epsilon^3)/2}$ we obtain:

$$\Pr\left[\|\mathbf{v}\| \geq (1+\epsilon)\|\mathbf{u}\|^2\right] \leq e^{-(\epsilon^2-\epsilon^3)k/4}$$

2.3 Bounding the other direction

$$\begin{aligned}
\Pr\left[\|\mathbf{v}\| \leq (1-\epsilon)\|\mathbf{u}\|^2\right] &= \Pr\left[x \leq (1-\epsilon)k\right] \\
&= \Pr\left[e^{-\lambda x} \geq e^{-\lambda(1-\epsilon)k}\right] \\
&= \left(\frac{E(e^{-\lambda x_1^2})}{e^{-\lambda(1-\epsilon)}}\right)^k \\
&\leq \left(\frac{e^{2(1-\epsilon)\lambda}}{1+2\lambda}\right)^{k/2} \qquad , \lambda = \frac{\epsilon}{2(1-\epsilon)} \\
&\leq ((1-\epsilon)e^{\epsilon})^{k/2} \\
&\leq e^{-(\epsilon^2-\epsilon^3)k/4}
\end{aligned}$$

Achieving for one point:

$$(1-\epsilon)\|\mathbf{u}\|^2 \leq \|\mathbf{v}\|^2 \leq (1+\epsilon)\|\mathbf{u}\|^2$$

with probability $\Pr \geq 1 - 2e^{-(\epsilon^2-\epsilon^3)\frac{k}{4}}$

2.4 Putting it all together

In the case we have n points we have to preserve $O(n^2)$ distances which are all the pairwise distances between points. We can fail with probability less then a constant, say, $\Pr_{fail} < 1/2$. We use a simple union bound:

$$\begin{aligned}
n^2 2e^{-(\epsilon^2-\epsilon^3)k/4} &< 1/2 \\
(\epsilon^2-\epsilon^3)k/4 &> 2\ln(2n) \\
k &> \frac{9\ln(n)}{\epsilon^2-\epsilon^3} \quad \text{For } n > 16.
\end{aligned}$$

Finally, since the failure probability is smaller then $1/2$ we can repeat until success a constant number of times in expectancy.

3 Fast low rank approximation

The results from the last section can be applied to accelerating low rank approximation of matrices. An optimal low rank approximations can be easily computed using the SVD of A in $O(mn^2)$. Using random projections we show how to achieve an "almost optimal" low rank approximation in $O(mn\log(n))$. We will go over a two step algorithm, suggested by Papadimitriou, Raghavan, Tamaki, and Vempala. First, we use k random projections to find a matrix B which is "much smaller" then A, but still shares most of its (right) eigenspace. Then, we SVD B and project A on B's k top eigenvectors.

3.1 Introduction

A low rank approximation of an $m \times n$, $(m \geq n)$, matrix A is another matrix A_k such that:

1. The rank of A_k is at most k.
2. $\|A - A_k\|_{norm}$ is minimized.

It is well known that for both l_2, and the Frobenius norms

$$A_k = \sum_{i=1}^{k} \sigma_i \mathbf{u}_i \mathbf{v}_i^T$$

where the singular value decomposition (SVD) of A is:

$$A = USV^T = \sum_{i=1}^{n} \sigma_i \mathbf{u}_i \mathbf{v}_i^T$$

The algorithm: Let R be an $m \times \ell$ matrix such that $R(i,j)$ are drawn i.i.d from $N(0,1)$. Also we have that $\ell \geq \frac{c \log(n)}{\epsilon^2}$.

1. Compute $B = \frac{1}{\sqrt{\ell}} R^T A$.
2. Compute the SVD of B, $B = \sum_{i=1}^{\ell} \lambda_i \mathbf{a}_i \mathbf{b}_i^T$.
3. Return: $\widetilde{A}_k \leftarrow A \left(\sum_{i=1}^{k} \mathbf{b}_i \mathbf{b}_i^T \right)$.

4 Proof of the algorithm

Since A_k is the optimal solution we cannot hope to do better then $\|A - A_k\|_F^2$, Yet we show that we do not do much worse.

$$\|A - \widetilde{A}_k\|_F^2 \leq \|A - A_k\|_F^2 + 2\epsilon \|A_k\|_F^2$$

Reminder:

$$A = \sum_{i=1}^{n} \sigma_i \mathbf{u}_i \mathbf{v}_i^T \quad , \quad A_k = \sum_{i=1}^{k} \sigma_i \mathbf{u}_i \mathbf{v}_i^T$$
$$B = \sum_{i=1}^{\ell} \lambda_i \mathbf{a}_i \mathbf{b}_i^T \quad , \quad \widetilde{A}_k = A \left(\sum_{i=1}^{k} \mathbf{b}_i \mathbf{b}_i^T \right).$$

Since the b_is are orthogonal

$$\begin{aligned}
\|A - \widetilde{A}_k\|_F^2 &= \sum_{i=1}^{n} \|(A - \widetilde{A}_k)\mathbf{b}_i\|^2 \\
&= \sum_{i=1}^{n} \|A\mathbf{b}_i - A(\sum_{j=1}^{k} \mathbf{b}_j \mathbf{b}_j^T)\mathbf{b}_i\|^2 \\
&= \sum_{i=k+1}^{n} \|A\mathbf{b}_i\|^2 \\
&= \|A\|_F^2 - \sum_{i=1}^{k} \|A\mathbf{b}_i\|^2
\end{aligned}$$

on the other hand:

$$\|A - A_k\|_F^2 = \|A\|_F^2 - \|A_k\|_F^2.$$

We gain that:

$$\|A - \widetilde{A}_k\|_F^2 = \|A - A_k\|_F^2 + (\|A_k\|_F^2 - \sum_{i=1}^{k} \|A\mathbf{b}_i\|^2).$$

We now need to show that $\|A_k\|_F^2$ is not much larger then $\sum_{i=1}^{k} \|Ab_i\|^2$. We start by giving a lower bound on $\sum_{i=1}^{k} \|Ab_i\|^2$.

$$\begin{aligned}
\sum_{i=1}^{k} \lambda_i^2 &= \sum_{i=1}^{k} \|B\mathbf{b}_i\|^2 \\
&= \sum_{i=1}^{k} \|\frac{1}{\sqrt{\ell}} R^T (A\mathbf{b}_i)\|^2 \\
&\leq (1+\epsilon) \sum_{i=1}^{k} \|Ab_i\|^2 \ \text{ With probability}
\end{aligned}$$

and so we gain

$$\sum_{i=1}^{k} \|Ab_i\|^2 \geq \frac{1}{1+\epsilon} \sum_{i=1}^{k} \lambda_i^2$$

We now need to bound the term $\sum_{i=1}^{k} \lambda_i^2$ from below:

$$\begin{aligned}
\sum_{i=1}^{k} \lambda_i^2 &\geq \sum_{i=1}^{k} \mathbf{v}_i^T B^T B \mathbf{v}_i \\
&= \sum_{i=1}^{k} \frac{1}{\ell} \mathbf{v}_i^T A^T R R^T A \mathbf{v}_i \\
&= \sum_{i=1}^{k} \sigma_i^2 \|\frac{1}{\sqrt{\ell}} R^T \mathbf{u}_i\|^2 \\
&\geq \sum_{i=1}^{k} (1-\epsilon)\sigma_i^2, \ \text{ With probability.}
\end{aligned}$$

Combining the inequalities with the fact that $\sum_{i=1}^{k} \sigma_i^2 = \|A_k\|_F^2$ we get:

$$\begin{aligned}
\sum_{i=1}^{k} \|Ab_i\|^2 &\geq \frac{1-\epsilon}{1+\epsilon} \|A_k\|_F^2 \\
&\geq (1-2\epsilon)\|A_k\|_F^2
\end{aligned}$$

Finally:

$$\|A - \widetilde{A}_k\|_F^2 \leq \|A - A_k\|_F^2 + 2\epsilon \|A_k\|_F^2$$

Notice that we need to preserve the length of at most $2n$ vectors. This gives us a success probability of $\Pr_{success} \geq 1 - 4ne^{-(\epsilon^2 - \epsilon^3)\ell/4}$, which is constant for $l = O(\frac{\log(n)}{\epsilon^2})$.

4.1 Computational savings for full matrices

1. Computing the matrix B is $O(mn\ell)$.
2. Computing the SVD of B is $O(m\ell^2)$.
3. Where $\ell \geq \frac{c\log(n)}{\epsilon^2}$.

Hence the total running time is:

$$O\left(\frac{1}{\epsilon^2} mn \log(n)\right).$$

as compared to the straightforward SVD which is $O(mn^2)$.

5 Summary

- We showed that a random matrix can be used to project points in $\mathbb{R}^d$ to $\mathbb{R}^k$ in a length preserving way (with high probability). Such that $k = O(\frac{\log(n)}{\epsilon^2})$.
- We used this fact to accelerate Low rank approximations of $m \times n$ matrices from $O(mn^2)$ to $O(mn\log(n))$.

6 Important message

Petros Drineas, Rafi Ostrovski and Yuval Rabani are organizing a research/reading group on algorithmic aspects of k-means clustering. The first meeting will take place this Friday, Sep 28, at 10:30am, at the IPAM lobby.

The topics are related and somewhat in the same spirit as the this talk. A list of the papers to be discussed during the semester were sent to you by email. The first paper read will be Kanungo et al.