

Random Projections

Instructors: Sham Kakade and Greg Shakhnarovich

1 The Johnson-Lindenstrauss lemma

Theorem 1.1. (*Johnson-Lindenstrauss*) Let $\epsilon \in (0, 1/2)$. Let $Q \subset \mathbb{R}^d$ be a set of n points and $k = \frac{20 \log n}{\epsilon^2}$. There exists a Lipschitz mapping $f : \mathbb{R}^d \rightarrow \mathbb{R}^k$ such that for all $u, v \in Q$:

$$(1 - \epsilon)\|u - v\|^2 \leq \|f(u) - f(v)\|^2 \leq (1 + \epsilon)\|u - v\|^2$$

To prove JL, we appeal to the following lemma:

Theorem 1.2. (*Norm preservation*) Let $x \in \mathbb{R}^d$. Assume that the entries in $A \subset \mathbb{R}^{k \times d}$ are sampled independently from $N(0, 1)$. Then,

$$\Pr((1 - \epsilon)\|x\|^2 \leq \|\frac{1}{\sqrt{k}}Ax\|^2 \leq (1 + \epsilon)\|x\|^2) \geq 1 - 2e^{-(\epsilon^2 - \epsilon^3)k/4}$$

The proof of the JL just appeals to the union bound:

Proof. The proof is constructive and is an example of the *probabilistic method*. Choose an f which is a *random projection*. Let $f = \frac{1}{\sqrt{k}}Ax$ where A is a $k \times d$ matrix, where each entry is sampled i.i.d from a Gaussian $N(0, 1)$.

Note there are $O(n^2)$ pairs of $u, v \in Q$. By the union bound,

$$\begin{aligned} & \Pr(\exists u, v \text{ s.t. the following event fails: } (1 - \epsilon)\|u - v\|^2 \leq \|\frac{1}{\sqrt{k}}A(u - v)\|^2 \leq (1 + \epsilon)\|u - v\|^2) \\ & \leq \sum_{u, v \in Q} \Pr(\text{ s.t. the following event fails: } (1 - \epsilon)\|u - v\|^2 \leq \|\frac{1}{\sqrt{k}}A(u - v)\|^2 \leq (1 + \epsilon)\|u - v\|^2) \\ & \leq 2n^2 e^{-(\epsilon^2 - \epsilon^3)k/4} \\ & < 1 \end{aligned}$$

the last step follows if we choose $k = \frac{20}{\epsilon^2} \log n$.

Note that the probability of finding a map f which satisfies the desired conditions is strictly greater than 0, so such a map must exist. (Aside: this proof technique is known as ‘the probabilistic method’ — note that the theorem is a deterministic statement while the proof is via a probabilistic argument.) \square

Now let us prove the norm preservation lemma:

Proof. First let us show that for any $x \in \mathbb{R}^d$, we have that:

$$\mathbb{E}[\|\frac{1}{\sqrt{k}}Ax\|^2] = \mathbb{E}[\|x\|^2].$$

To see this, let us examine the expected value of the entry $[Ax]_j^2$

$$\begin{aligned}
\mathbb{E}[[Ax]_j^2] &= \mathbb{E}\left[\left(\sum_{i=1}^d A_{i,j}x_i\right)^2\right] \\
&= \mathbb{E}\left[\sum_{i,i'} A_{i,j}A_{i',j}x_i x_{i'}\right] \\
&= \mathbb{E}\left[\sum_i A_{i,i}^2 x_i^2\right] \\
&= \sum_i x_i^2 \\
&= \|x\|^2
\end{aligned}$$

and note that:

$$\left\|\frac{1}{\sqrt{k}}Ax\right\|^2 = \frac{1}{k}\sum_{j=1}^k [Ax]_j^2$$

which proves the first claim (note that all we require for this proof is independence and unit variance in constructing A).

Note that above shows that $\tilde{Z}_j = [Ax]_j/\|x\|$ is distributed as $N(0, 1)$, and \tilde{Z}_j are independent. We now bound the failure probability of one side. By the union bound,

$$\begin{aligned}
\Pr\left(\left\|\frac{1}{\sqrt{k}}Ax\right\|^2 > (1 + \epsilon)\|x\|^2\right) &= \Pr\left(\sum_{i=1}^k \tilde{Z}_i^2 > (1 + \epsilon)k\right) \\
&= n^2 \Pr(\chi_k^2 > (1 + \epsilon)k)
\end{aligned}$$

(where χ_k^2 is the chi-squared distribution with k degrees of freedom). Now we appeal to a concentration result below, which bounds this probability by:

$$\leq \exp\left(-\frac{k}{4}(\epsilon^2 - \epsilon^3)\right)$$

A similar argument handles the other side (and the factor of 2 in the bound). □

The following lemma for χ^2 - distributions was used in the above proof.

Lemma 1.3. *We have that:*

$$\begin{aligned}
\Pr(\chi_k^2 \geq (1 + \epsilon)k) &\leq \exp\left(-\frac{k}{4}(\epsilon^2 - \epsilon^3)\right) \\
\Pr(\chi_k^2 \leq (1 - \epsilon)k) &\leq \exp\left(-\frac{k}{4}(\epsilon^2 - \epsilon^3)\right)
\end{aligned}$$

Proof. Let Z_1, Z_2, \dots, Z_k be i.i.d. $N(0, 1)$ random variables. By Markov's inequality,

$$\begin{aligned}
\Pr(\chi_k^2 \geq (1 + \epsilon)k) &= \Pr\left(\sum_{i=1}^k Z_i^2 > (1 + \epsilon)k\right) \\
&= \Pr\left(e^{\lambda \sum_{i=1}^k Z_i^2} > e^{(1+\epsilon)k\lambda}\right) \\
&\leq \frac{\mathbb{E}[e^{\lambda \sum_{i=1}^k Z_i^2}]}{e^{(1+\epsilon)k\lambda}} \\
&= \frac{(\mathbb{E}[e^{\lambda Z_1^2}])^k}{e^{(1+\epsilon)k\lambda}} \\
&= e^{-(1+\epsilon)k\lambda} \left(\frac{1}{1-2\lambda}\right)^{k/2}
\end{aligned}$$

where the last step follows from evaluating the expectation, which holds for $0 < \lambda \leq 1/2$ (this expectation is just the moment generating function). Choosing $\lambda = \frac{\epsilon}{2(1+\epsilon)}$ which minimizes the above expression (and is less than $1/2$ as required), we have:

$$\begin{aligned}
\Pr(\chi_k^2 \geq (1 + \epsilon)k) &= ((1 + \epsilon)e^{-\epsilon})^{\frac{k}{2}} \\
&\leq \exp\left(-\frac{k}{4}(\epsilon^2 - \epsilon^3)\right)
\end{aligned}$$

using the upper bound $1 + \epsilon \leq \exp(\epsilon - (\epsilon^2 - \epsilon^3)/2)$. The other bound is proved in a similar manner. \square

The following lemma shows that nothing is fundamental about using Gaussian in particular. Many distributions with unit variance and certain boundedness properties (or higher order moment conditions) suffice.

Lemma 1.4. *Assume for $A \in \mathbb{R}^{k \times d}$ that each A_i, j is uniform on $\{-1, 1\}$. Then for any vector $x \in \mathbb{R}^d$:*

$$\begin{aligned}
\Pr\left(\left\|\frac{1}{\sqrt{k}}Ax\right\|^2 \geq (1 + \epsilon)\|x\|^2\right) &\leq \exp\left(-\frac{k}{4}(\epsilon^2 - \epsilon^3)\right) \\
\Pr\left(\left\|\frac{1}{\sqrt{k}}Ax\right\|^2 \leq (1 - \epsilon)\|x\|^2\right) &\leq \exp\left(-\frac{k}{4}(\epsilon^2 - \epsilon^3)\right)
\end{aligned}$$

2 Preserving Inner Products

As a simple corollary, we see that inner products are preserved under random projection.

Corollary 2.1. *Let $u, v \in \mathbb{R}^d$ and that $\|u\| \leq 1$ and $\|v\| \leq 1$. Let $f = \frac{1}{\sqrt{k}}Ax$ where A is a $k \times d$ matrix, where each entry is sampled i.i.d from a Gaussian $N(0, 1)$ (or from $U(-1, 1)$). Then,*

$$\Pr(|u \cdot v - f(u) \cdot f(v)| \geq \epsilon) \leq 4e^{-(\epsilon^2 - \epsilon^3)k/4}$$

Proof. Applying Theorem 1.2 to the vectors $u + v$ and $u - v$, we have that with probability at least $1 - 4e^{-(\epsilon^2 - \epsilon^3)k/4}$:

$$\begin{aligned}
(1 - \epsilon)\|u + v\|^2 &\leq \|f(u + v)\|^2 \leq (1 + \epsilon)\|u + v\|^2 \\
(1 - \epsilon)\|u - v\|^2 &\leq \|f(u - v)\|^2 \leq (1 + \epsilon)\|u - v\|^2
\end{aligned}$$

Now we have:

$$\begin{aligned}
4f(u) \cdot f(v) &= \|f(u + v)\|^2 - \|f(u - v)\|^2 \\
&\geq (1 - \epsilon)\|u + v\|^2 - (1 + \epsilon)\|u - v\|^2 \\
&= 4u \cdot v - 2\epsilon(\|u\|^2 + \|v\|^2) \\
&\geq 4u \cdot v - 4\epsilon
\end{aligned}$$

The proof of the other direction is analogous. \square

3 References

Material used was from Santosh Vempala's monograph on random projections.