Topics in High-Dimensional Probability and Statistics*

Lecture 5: Random projections and the Johnson-Lindenstrauss lemma

Contents

1 Approximate isometries

Consider n distinct data points $x_1, \ldots, x_n$ in $\mathbb{R}^D$ considered deterministic (all the following results may be easily extended to the case of random points via conditioning). If the dimension D is very large, processing this data for some given task may be computationally demanding. An interesting problem is to figure out whether there exists a way to transform the high-dimensional data points $x_1, \ldots, x_n \in \mathbb{R}^D$, through some map

$$T : \mathbb{R}^D \to \mathbb{R}^d \quad \text{for some} \quad d \ll D,$$

into lower dimensional data points $T(x_1), \ldots, T(x_n) \in \mathbb{R}^d$ without losing too much information about the original data.

One way to guarantee that map T preserves the information of the data is to require the geometry of the data set to be completely preserved, i.e., to require that $T : \{x_1, \ldots, x_n\} \to \mathbb{R}^d$ is an isometry. Precisely, this means that, for all $i \neq j$,

$$\|T(x_i) - T(x_j)\|_2 = \|x_i - x_j\|_2,$$

where, on the left hand-side, $\|.\|_2$ refers to the euclidean norm in $\mathbb{R}^d$ while, on the right hand-side, $\|.\|_2$ refers to the euclidean norm in $\mathbb{R}^D$.

This isn't really a reasonable requirement for many reasons. First, one can exhibit simple settings in which it is impossible when we restrict attention to linear maps (see example 1.1).

Example 1.1. *Consider $D = 2$, $d = 1$ and $x_1, x_2, x_3 \in \mathbb{R}^2$ the vertices of a triangle with sides of equal lengths. Then, there is no linear map $T : \mathbb{R}^2 \to \mathbb{R}$ that preserves pairwise distances.*

More generally, if we think of the data points as points sampled from a distribution with a density with respect to Lebesgue measure, then for any $d < D$, the points $x_1, \ldots, x_n$ all belong to a strict subspace of $\mathbb{R}^D$ (i.e., a subspace of dimension at most $D - 1$) with probability 0. Hence, mapping all these points isometrically into a lower dimensional space is likely to fail with high probability.

But one can be a little less demanding, and require T to be an approximate isometry. To be more precise, for a fixed $\varepsilon \in (0, 1)$, we could only ask to have, for all $i \neq j$,

$$1 - \varepsilon \leq \frac{\|T(x_i) - T(x_j)\|_2^2}{\|x_i - x_j\|_2^2} \leq 1 + \varepsilon.$$

The goal of this lecture is to show that we can construct a random and linear map $T : \mathbb{R}^D \to \mathbb{R}^d$ such that, for any every $\varepsilon, \delta \in (0, 1)$, the above property holds with probability $1 - \delta$ for d of order

$$\frac{1}{\varepsilon^2} \log\left(\frac{n}{\sqrt{\delta}}\right),$$

and independently of the dimension D.

2 Reminder

We recall a few facts, seen in lecture 2, that will be useful in the proof of the Johnson-Lindenstrauss lemma below.

A basic result of interest will be the following simple version of the Bernstein's concentration inequality.

Lemma 2.1. *Let $Y_1, \ldots, Y_n$ be independent random variables. Suppose that there exists $s^2, b > 0$ such that, for all $1 \leq i \leq n$ and for all $\theta \in [-1/b, 1/b]$,*

$$\log \mathbb{E} \exp(\theta\{Y_i - \mathbb{E}Y_i\}) \leq \frac{\theta^2 s^2}{2}.$$

Then, for all $t > 0$,

$$\mathbb{P}\left\{\frac{1}{n}\sum_{i=1}^n (Y_i - \mathbb{E}Y_i) > t\right\} \vee \mathbb{P}\left\{\frac{1}{n}\sum_{i=1}^n (Y_i - \mathbb{E}Y_i) < -t\right\} \leq \exp\left(-\frac{nt}{2}\left\{\frac{1}{b} \wedge \frac{t}{s^2}\right\}\right).$$

The second important observation is that, given a real valued and sub-gaussian random variable X with variance proxy σ^2, the variable X^2 satisfies,

$$\forall \theta \in (-\frac{1}{a}, \frac{1}{a}), \quad \log \mathbb{E}[\exp(\theta\{X^2 - \mathbb{E}X^2\})] \leq \frac{\theta^2 a^2}{2(1 - \theta a)},$$

with

$$a := 4e\sigma^2.$$

In particular,

$$\forall \theta \in [-\frac{1}{2a}, \frac{1}{2a}], \quad \log \mathbb{E}[\exp(\theta\{X^2 - \mathbb{E}X^2\})] \leq \frac{\theta^2 (2a^2)}{2}.$$

3 Johnson-Lindenstrauss lemma

Let $\mathcal{X} = \{x_1, \ldots, x_n\} \subset \mathbb{R}^D$ be a set of n distinct data points, considered deterministic, and fix

$$\varepsilon, \delta \in (0, 1).$$

*Teaching material can be found at https://www.qparis-math.com/teaching.

Theorem 3.1. *Let $M \in \mathbb{R}^{d\times D}$ be a random matrix whose rows $R_1, \ldots, R_d \in \mathbb{R}^D$ are independent, centered and isotropic, i.e., such that*

$$\mathbb{E}[R_i] = 0 \quad \textit{and} \quad \mathbb{E}[R_i R_i^\top] = I_D.$$

Suppose that each R_i is sub-gaussian with variance proxy at most σ^2. Define finally

$$T := \frac{1}{\sqrt{d}} M.$$

Then, provided

$$d \geq \frac{64e^2\sigma^4}{\varepsilon^2} \log\left(\frac{n^2}{\delta}\right),$$

we have

$$\mathbb{P}\left(\forall i \neq j : 1 - \varepsilon \leq \frac{\|T(x_i) - T(x_j)\|_2^2}{\|x_i - x_j\|_2^2} \leq 1 + \varepsilon\right) \geq 1 - \delta.$$

Proof. Denote

$$\mathcal{Z} := \left\{\frac{x_i - x_j}{\|x_i - x_j\|_2} : i \neq j\right\}.$$

By linearity of T, the statement we need to prove is then equivalent to

$$\mathbb{P}\left(\max_{z\in\mathcal{Z}} |\|T(z)\|_2^2 - 1| > \varepsilon\right) < \delta.$$

Using a union bound, observe that

$$\begin{aligned}
&\mathbb{P}\left(\max_{z\in\mathcal{Z}} |\|T(z)\|_2^2 - 1| > \varepsilon\right) \\
&\leq |\mathcal{Z}| \max_{z\in\mathcal{Z}} \mathbb{P}\left(|\|T(z)\|_2^2 - 1| > \varepsilon\right) \\
&= \frac{n(n-1)}{2} \max_{z\in\mathcal{Z}} \mathbb{P}\left(|\|T(z)\|_2^2 - 1| > \varepsilon\right) \\
&< \frac{n^2}{2} \max_{z\in\mathcal{Z}} \mathbb{P}\left(|\|T(z)\|_2^2 - 1| > \varepsilon\right).
\end{aligned}$$

As a result, it is enough to prove that, for all $z \in \mathcal{Z}$,

$$\mathbb{P}\left(|\|T(z)\|_2^2 - 1| > \varepsilon\right) \leq \frac{2\delta}{n^2}.$$

For $z \in \mathcal{Z}$, note that

$$\begin{aligned}
T(z) &= \frac{1}{\sqrt{d}} Mz \\
&= \frac{1}{\sqrt{d}} (\langle R_1, z\rangle, \ldots, \langle R_d, z\rangle)^\top.
\end{aligned}$$

As a result,

$$\begin{aligned}
|\|T(z)\|_2^2 - 1| &= |\frac{1}{d}\sum_{i=1}^d \langle R_i, z\rangle^2 - 1| \\
&= |\frac{1}{d}\sum_{i=1}^d (\langle R_i, z\rangle^2 - \mathbb{E}\langle R_i, z\rangle^2)|,
\end{aligned}$$

where the last identity follows since

$$\mathbb{E}[\langle R_i, z\rangle^2] = z^\top \mathbb{E}[R_i R_i^\top] z = \|z\|_2^2 = 1.$$

Since $\|z\|_2 = 1$ for every $z \in \mathcal{Z}$, each random variable $\langle R_i, z\rangle$ is sub-gaussian with variance proxy at most σ^2. According to results mentioned in the previous section, this implies that variables

$$Y_i := \langle R_i, z\rangle^2,$$

satisfy, for all $1 \leq i \leq d$ and for all $\theta \in [-1/b, 1/b]$,

$$\log \mathbb{E}\exp(\theta\{Y_i - \mathbb{E}Y_i\}) \leq \frac{\theta^2 s^2}{2},$$

where $b = 8e\sigma^2$ and $s^2 = 32e^2\sigma^4$. As a result, we deduce that, for every $z \in \mathcal{Z}$,

$$\begin{aligned}
\mathbb{P}\left(|\|T(z)\|_2^2 - 1| > \varepsilon\right) &\leq 2\exp\left(-\frac{d\varepsilon}{2}\left\{\frac{1}{b} \wedge \frac{\varepsilon}{s^2}\right\}\right) \\
&= 2\exp\left(-\frac{d\varepsilon}{16e\sigma^2}\left\{1 \wedge \frac{\varepsilon}{4e\sigma^2}\right\}\right) \\
&= 2\exp\left(-\frac{d\varepsilon^2}{64e^2\sigma^4}\right),
\end{aligned}$$

where the last inequality follows from the fact that $\varepsilon \in (0, 1)$ and that $\sigma^2 \geq 1/4e$ by assumption. To sum up, the statement follows provided

$$2\exp\left(-\frac{d\varepsilon^2}{64e^2\sigma^4}\right) \leq \frac{2\delta}{n^2},$$

which is equivalent to

$$d \geq \frac{64e^2\sigma^4}{\varepsilon^2} \log\left(\frac{n^2}{\delta}\right).$$

□

4 Examples

We give two explicit constructions of matrix M satisfying the assumptions of the theorem.

Example 4.1. *Suppose that $M = (M_{i,j})$ where entries $M_{i,j}$ are independent and, for all $i \in \{1, \ldots, d\}$ and all $j \in \{1, \ldots, D\}$,*

$$\mathbb{P}(M_{i,j} = -1) = \mathbb{P}(M_{i,j} = +1) = \frac{1}{2}.$$

Then it satisfies the assumptions of Theorem 3.1 with $\sigma^2 = 1$.

Example 4.2. *Suppose that $M = (M_{i,j})$ where entries $M_{i,j}$ are independent and, for all $i \in \{1, \ldots, d\}$ and all $j \in \{1, \ldots, D\}$,*

$$M_{i,j} \sim \mathcal{N}(0, 1).$$

Then it satisfies the assumptions of Theorem 3.1 with $\sigma^2 = 1$.

5 Note

For an application of Theorem 3.1 in the context of clustering, we refer the reader to [2]. We also recommend Chapter 5 in [1] for further applications of the Johnson-Lindenstrauss lemma.

References

[1] A. Bandeira. Ten lectures and forty-two open problems in the mathematics of data science. Lecture notes, 2016.

[2] G. Biau, L. Devroye, and G. Lugosi. On the performance of clustering in Hilbert spaces. *IEEE Trans. Inform. Theory*, 54(2):781–790, 2008.