Lecture 6: September 17

Lecturer: Alessandro Rinaldo *Scribes: Sangwon Hyun*

Note: *LaTeX template courtesy of UC Berkeley EECS dept.*

Disclaimer: *These notes have not been subjected to the usual scrutiny reserved for formal publications. They may be distributed outside this class only with the permission of the Instructor.*

This lecture's notes illustrate some uses of various LaTeX macros. Take a look at this and imitate.

6.1 Last Time: introduced Johnson-Lindenstrauss

We discussed Johnson-Lindenstrauss theorem for L_2 (for L_1, it doesn't work!), which basically states that an embedding of a set $s \subseteq \mathbb{R}^D$, $|S| = n$ into $\mathbb{R}^d$ so that pairwise distances are approximated up to $1 \pm \epsilon$ factor, $\epsilon \in (0, 1)$, for $D >> \Phi$ and $d = O(\epsilon^{-2} \log n)$. In English: "It is possible to find random lower-dimensional representation of a high(D)-dimensional object in a smaller space.

6.2 Continuing with J-L

Dasgupta-Gupta has an elementary proof of Johnson-Lindenstrauss, regarding random projections:

Theorem 6.1 *For all $\epsilon \in (0, 1)$ and positive integer n, let $d \geq 4\left(\frac{\epsilon^2}{2} + \frac{\epsilon^3}{3}\right)^{-1} \log n$. Then, for any set $S \subseteq \mathbb{R}^D$ of n points, there exists a map $f : \mathbb{R}^D \to \mathbb{R}^d$ such that for all $x, y \in S$, the following holds:*

$$(1 - \epsilon)\|x - y\|^2 \leq \|f(x) - f(y)\|^2 \leq (1 + \epsilon)\|x - y\|^2 \tag{6.1}$$

Furthermore, f can be compared in polynomial time; in fact, it is a random map(projection) so that (6.1) hold with probability at least $1 - \frac{1}{n}$. The algorithm is to simulate a d-dimensional subspace of $\mathbb{R}^D$, spanned by d standard Gaussian vectors in $\mathbb{R}^D$ (with probability 1 they are linearly independent). [1]

Here is a computationally better method that approximates the random projection, due to Achlioptas (2001) [2]

1. Construct a matrix $W_{d\times D}$ whose entries are $\frac{1}{\sqrt{d}} X_{ij}$, where expectation and variance of independent sub-Gaussian X_{ij}'s each of which are $E[X_{ij}] = 0$ and $Var[X_{ij}] = 1$. (e.g. Radamacher, or ± 1)
2. Pick $\alpha \in \mathbb{R}^D$. Let $W_i(\alpha) = \sum_{j=1}^{D} \alpha_j X_{ij}$
3. $W(\alpha) = \frac{1}{\sqrt{d}} \begin{bmatrix} W_1(\alpha) \\ \vdots \\ W_d(\alpha) \end{bmatrix} = W_\alpha$.

[1] See chapter 14-15 of "Lectures in Geometry" if interested in this.
[2] Note, this is a way to get the approximate k-nearest-neighbor in high dimensions

Now we can see that $E[W_i(\alpha)^2] = \sum_{i=1}^{D} \alpha_j^2 = \|\alpha\|_2$, and $E[\|W_i(\alpha)\|^2] = \frac{1}{\alpha}\sum_{i=1}^{d} E[W_i^2(\alpha)] = \|\alpha\|_2$. We also want to show that:

$$(1-\epsilon)\|a\|^2 \leq \|W(\alpha)\|^2 \leq (1+\epsilon)\|\alpha\|^2,\ \forall \alpha \in \mathbb{R}^D,\ \text{with high probability}$$

We invoke a theorem that directly states this:

Theorem 6.2 *Let $S \subseteq \mathbb{R}^D$, $|s| = n$. Let X_{ij} be zero mean unit variance sub-Gaussian vectors. Also let $\delta, \epsilon \in (0,1)$. Then, if $d \geq \frac{100v^2}{\epsilon^2}\log\left(\frac{n}{\sqrt{\delta}}\right)$, W is an ϵ-isometry on S with probability $1-\alpha$. In other words,*

$$(1-\epsilon)\|x-y\|^2 \leq \|W(x-y)\|^2 \leq (1+\epsilon)\|x-y\|^2, \forall x, y \in S$$

Proof: Let $T = \{\frac{x-y}{\|x-y\|} : x, y \in S, x \neq y\} \subseteq \mathbb{S}^{D-1}$, $\|T\| \leq \binom{n}{2}$

We will show:

$$\max_{\alpha \in T}\left|\|W(\alpha)\|^2 - 1\right| \leq \epsilon$$

For $X_{ij} \in G(\nu)$, $\forall \alpha \in T$, the expectation of $e^{\lambda W_i(\alpha_j)}$ can be bounded as:

$$E[d^{\lambda W_i(\alpha)}] = \prod_{j=1}^{D} E[e^{\lambda \alpha_j x_j}] \leq \exp\{\frac{\lambda^2 y^2}{2}\sum_{j=1}^{D}\alpha_j^2\} = \exp\{\frac{\lambda^2\nu^2}{2}\}$$

from which we know that $W_i(\alpha) \in G(\nu)$ is sub-Gaussian for $\alpha \in T$. Now, we'll use a fun fact about sub-Gaussianity: If $X \in G(\nu)$, then $E[X^{2q}] \leq 2q!(2\nu)^q$.[3] Using this,

$$E[W_i(\alpha)^{2q}] \leq 2q!(2\nu)^q \leq \frac{q!}{2}(4\nu)^q, q = 2, 3, \cdots$$

Then, by Bernstein inequality (the general version, with $d(4\nu)^2$ instead of ν^2, 4ν instead of c)

$$\forall \alpha \in T,\ \ x > 0,\ \ \mathbb{P}(|\sum_{i=1}^{\alpha}(W_i(\alpha)^2 - 1)| \geq 4\nu\sqrt{2dx} + 4\nu x) \leq 2e^{-x}$$

and take union bound over T to get:

$$\mathbb{P}(|\max_{\alpha \in T}(W_i(\alpha)^2 - 1)| \geq \overbrace{d\epsilon}^{4\nu\sqrt{2dx}+4\nu x}) \leq \underbrace{|T|}_{\text{typically} \leq \binom{n}{2}} 2e^{-x} \leq n^2 e^{-x}$$

Set $x = \log\left(\frac{n^2}{\delta}\right)$ to get

$$\mathbb{P}\left(\left|\max_{\alpha \in T}\|W_i(\alpha)\|^2 - 1\right| > \underbrace{8\nu\sqrt{\frac{\log\frac{n}{\sqrt{\delta}}}{\alpha}} + \frac{8\nu\log\frac{n}{\sqrt{\delta}}}{\alpha}}_{\Delta}\right) \leq \delta$$

So, for $d \geq \frac{100}{\epsilon^2}\nu^2\log\left(\frac{n}{\sqrt{\delta}}\right)$, we can have

$$\Delta \leq \frac{4\epsilon}{5} + \frac{2}{25}\frac{\epsilon^2}{\nu} \leq \epsilon$$

[3] Proof starts with $E[X^{2q}] \leq \int_{-\infty}^{+infty} P(X^{2q} > t)dt$, and proceed $\cdots$

(because $\nu \geq 1$). ■

Note that this holds for any n points S, and the rate $O\left(\frac{\log n}{\epsilon^2}\right)$ is tight, unless you assume additional structure. [4]

Several applications are possible: KNN, hashing, compressed sensing etc. Furthermore, Achlioptas shows that bounds can be improved with some assumptions! If the mgf of $W_n(\alpha)$ is bounded by the mgf of χ_1^2 (which is $\frac{1}{\sqrt{1-2\lambda}}, \lambda < \frac{1}{2}$), then

$$\mathbb{P}\left(\left|\|W(\alpha\|^2 - \|\alpha\|^2\right| \geq \epsilon\|\alpha\|^2\right) \leq 2e^{-nC_\epsilon}$$

where $C_\epsilon = \frac{\epsilon^2}{2} - \frac{\epsilon^3}{\sigma}$. Gaussian, Radamacher $X_{ij} = \sqrt{3}\begin{cases} 1 & \text{with probability } 1/6 \\ 0 & \text{with probability } 2/3 \\ -1 & \text{with probability } 1/6 \end{cases}$ all work here.

6.3 Next topic: Bounding variance of Functions of Independent Random Variables

We introduce a useful inequality called Efron-Stein. Denote $X_1, \cdots, X_n$ independent, $f : \mathbb{R}^d \to \mathbb{R}$, and $Z = f(X_1, \cdots, X_n)$ whose second moment is finite $\mathbb{E}[Z^2] < \infty$. [5] The task is to bound the variance of Z. Also using notation $\mathbb{E}_i[]$ for the conditional expectation given $X_1, \cdots, X_i$ and $\mathbb{E}.[\cdot] = \mathbb{E}[\cdot]$ (the trivial sigma field) so that $E_n[Z] = Z$. Now, define $\Delta_i = \mathbb{E}_i[Z] - \mathbb{E}_{i-1}[Z]$ so that $Z - \mathbb{E}[Z] = \sum_{i=1}^n \Delta_i$, and $V[Z] = \sum_{i=1}^n \mathbb{E}[\Delta_i^2] + 2\sum_{i<j}\mathbb{E}[\Delta_i\Delta_j]$. If $j > i$, notice that $\mathbb{E}_i[\Delta_j] = \mathbb{E}_i[\mathbb{E}_j[\Delta_j]] - \mathbb{E}_i[\mathbb{E}_{i-1}[Z]] = \mathbb{E}_i[Z] - \mathbb{E}_i[Z] = 0$, by tower property of conditional expectation. So, if $j > i$, we have that $\mathbb{E}[\Delta_i\Delta_j] = \mathbb{E}[\mathbb{E}_i[\Delta_i\Delta_j]] = \mathbb{E}[\Delta_i \underbrace{\mathbb{E}_i[\Delta_j]}_{0}] = 0$, so that $V[Z] = \sum_{i=1}^n \mathbb{E}[\Delta_i^2]$. Also, lastly, denote $\mathbb{E}^{(i)}[]$ to be the conditional expectation given $X^{(i)} = \{X_j, j \neq i\}$, then $\mathbb{E}^{(i)}[Z] = \mathbb{E}^{(i)}[f(X_1, \ldots, X_n)] = \int f(X_1, \cdots, \underbrace{x}_{\text{i'th variable}}, \cdots, X_n)dP_i(x)$ and $\mathbb{E}_i[\mathbb{E}^{(i)}[Z]]$ (Phew!!) Now, we introduce the Efron Stein inequality which bounds the variance of Z.

Theorem 6.3 (Efron-Stein) *Let $X_1, \cdots, X_n$ be independent RVs and $Z = f(X_1, \cdots, X_n)$ be square integrable; then,*

$$V[Z] \leq \sum_{i=1}^n \mathbb{E}[(Z - \mathbb{E}^{(i)}(Z))^2] = \nu^2$$

Several alternative formulations are possible: if $X_1', \ldots, X_n'$ are independent copies of $X_1, \ldots, X_n$, and denote $Z_i' = f(X_1, \ldots, X_i', \ldots, X_n), i = 1, \cdots, n$. Then,

$$\begin{aligned}
\nu^2 &= \frac{1}{2}\sum_{i=1}^n \mathbb{E}[(Z - Z_i')^2] \\
&= \sum_{i=1}^n \mathbb{E}[[(Z - Z_i')_+]^2] \\
&= \sum_{i=1}^n \mathbb{E}[[(Z - Z_i')_-]^2]
\end{aligned}$$

[4] From this, we can sort of say that compressed sensing can be viewed as just an application of this Johnson-Lindenstrauss Lemma!

[5] Note, E-S was originally developed as research about jackknifed residuals for estimating bias.

and $\nu^2 = \inf_{Z_i} \sum_{i=1}^n \mathbb{E}[(Z - Z_i)^2]$, the infimum over all measureable functions of $X^{(i)}$. These are *all* equivalent representations!!

See Lecture 6, Thu Sep 17 of `http://www.stat.cmu.edu/~arinaldo/36788/references.html` for references and reading!