CS 229r: Algorithms for Big Data Fall 2013

Lecture 10 — October 3, 2013

Prof. Jelani Nelson *Scribe: Yong Wook Kwon*

1 Overview

In the last lecture, we have used the concentration of Lipschitz functions of Gaussians and the decoupling lemma to prove Hanson-Wright inequality, which subsequently implied the distributional JL lemma, which then finally implied the Johnson-Lindenstrauss lemma.

The goal of this lecture is as follows:

- Prove Lipschitz-concentration and Decoupling lemma
- Prove Alon's lower-bound ($m \gtrsim \Omega(\frac{1}{\varepsilon^2 \log \frac{1}{\varepsilon}} \log N)$)
- Try to circumvent Alon's lower-bound by giving a more refined upper bound for JL which takes properties of the point set into account,

2 Proof of Lipschitz Concentration

We shall prove a slightly more general theorem that implies Lipschitz concentration. The theorem was first stated by Pisier [1], but a more elegant proof, which is reproduced below, was later given by Maurey.

Theorem 1. *Let $X = (x_1, x_2, ...x_n)$ be i.i.d. $N(0,1)$, $\Phi : \mathbb{R} \mapsto \mathbb{R}$ convex, and $f : \mathbb{R}^n \mapsto \mathbb{R}$ has a gradient almost everywhere (except a set of measure zero). Then, $\mathbb{E}_X \Phi(f(x) - \mathbb{E} f(X)) \leq \mathbb{E}_{X,Y} \Phi(\frac{\pi}{2} \langle \nabla f(X), Y \rangle)$, where $Y = (y_1, ...y_n)$ also has i.i.d. $N(0,1)$ entries.*

Note that this theorem implies Lipschitz concentration, if one chooses $\Phi(z) = |z|^p$. Then,

$$\mathop{\mathbb{E}}_X |f(x) - \mathbb{E} f(X)|^p \leq (\frac{\pi}{2})^p \mathop{\mathbb{E}}_X \mathop{\mathbb{E}}_Y |\langle \nabla f(x), y \rangle|^p$$

As for a fixed X, the term inside the expectation is a convolution of normal variables, so by the 2-stableness of the normal, has distribution $g \cdot \|\nabla f(X)\|_2$. Then, we use that the p-th moment of the normal is bounded upwards by $O(\sqrt{p})^p$ to obtain

$$(\frac{\pi}{2})^p \mathop{\mathbb{E}}_X \mathop{\mathbb{E}}_Y |\langle \nabla f(x), y \rangle|^p \leq (\frac{c\pi}{2})^p (\sqrt{p})^p \mathop{\mathbb{E}}_X \|\nabla f(x)\|_2^p$$

But note that the ℓ_2 norm of the gradient is at most the Lipschitz constant, since by moving an infinitesimal amount ε (in ℓ_2 distance) in the direction of the gradient, we change f by ε times the

gradient. Thus, taking the $\frac{1}{p}$th power of both sides of the inequality gives Lipschitz Concentration, as desired.

Let us now prove the original theorem.

Proof.

$$\mathop{\mathbb{E}}_{X} \Phi(f(X) - \mathbb{E}\, f(X)) = \mathop{\mathbb{E}}_{X} \Phi(\mathop{\mathbb{E}}_{Y}(f(X) - f(Y))) \leq \mathop{\mathbb{E}}_{X,Y} \Phi(f(X) - f(Y)) \text{ (Jensen)}$$

This technique of pulling out the expectation out of the convex function via Jensen and making the inner expression symmetric in terms of X and Y is a useful technique and usually called *symmetrization*.

But the proof only gets slicker from here.

Define $g(\theta) = X\sin(\theta) + Y\cos(\theta)$, and note $\frac{d}{d\theta}g(\theta) = X\cos(\theta) - Y\sin(\theta) \stackrel{\text{def}}{=} g'(\theta)$.

$$\begin{aligned}
\mathop{\mathbb{E}}_{X,Y} \Phi(f(x) - f(Y)) &= \mathop{\mathbb{E}}_{X,Y} \Phi(f(g(\frac{\pi}{2})) - f(g(0))) \\
&= \mathop{\mathbb{E}}_{X,Y} \Phi(\int_0^{\frac{\pi}{2}} \frac{d}{d\theta}(f(g(\theta)))d\theta) \text{ (fundamental theorem of calculus)} \\
&= \mathop{\mathbb{E}}_{X,Y} \Phi(\int_0^{\frac{\pi}{2}} \left\langle \nabla f(g(\theta)), g'(\theta) \right\rangle d\theta) \\
&= \mathop{\mathbb{E}}_{X,Y} \Phi(\int_0^{\frac{\pi}{2}} \frac{2}{\pi}[\frac{\pi}{2} \left\langle \nabla f(g(\theta)), g'(\theta) \right\rangle]d\theta) \\
&= \mathop{\mathbb{E}}_{X,Y} \Phi(\mathop{\mathbb{E}}_{\theta\in[0,\frac{\pi}{2}]}(\frac{\pi}{2} \left\langle \nabla f(g(\theta)), g'(\theta) \right\rangle)) \\
&\leq \mathop{\mathbb{E}}_{X,Y,\theta} \Phi(\frac{\pi}{2} \left\langle \nabla f(g(\theta)), g'(\theta) \right\rangle) \text{ (Jensen)} \\
&= \mathop{\mathbb{E}}_{X,Y} \Phi(\frac{\pi}{2} \left\langle \nabla f(X), Y \right\rangle)
\end{aligned}$$

which completes the proof. The last inequality holds because for any fixed θ, $[g'(\theta), g(\theta)']^T$ is formed by applying a two by two rotation matrix to $[X, Y]$. Thus since X, Y are distributed as independent gaussians, then so are $g'(\theta), g(\theta)$ (drawing the point (X, Y) in $\mathbb{R}^2$, we see that choosing two independent gaussians is equivalent to choosing a uniformly random point on a circle of radius r, where r is chosen from the appropriate distribution; applying some fixed rotation still gives two independent and uniformly random points on the circle). □

3 Proof of De-coupling

Recall the statement from [4]: $\sigma_1, \sigma_1', \ldots, \sigma_n, \sigma_n'$ are i.i.d. signs, $A = (a_{ij})$, then

$$\|\sum_{i\neq j} a_{ij}\sigma_i\sigma_j\|_p \leq 4\|\sum_{i,j} a_{ij}\sigma_i\sigma_j'\|_p$$

"There are going to be lots of random variables in this proof" - Jelani Nelson

Note that the p-norm is a norm in the sense of random variables, that is $(\mathbb{E}\,|X|^p)^{\frac{1}{p}}$. We will use the trick of inserting new random variables and pulling them out, in an opportune moment.

Proof. Let $y_1, ...y_n \in \{0,1\}$ be fair coin flips. Then,

$$\|\sum_{i\neq j} a_{ij}\sigma_i\sigma_j\|_p = 4\|\mathop{\mathbb{E}}_{y}\sum_{i\neq j} a_{ij}\sigma_i y_i\sigma_j(1-y_j)\| \leq 4\|\sum_{j\neq i} a_{ij}\sigma_i y_i\sigma_j(1-y_j)\|_p \text{ (Jensen)}$$

Note that as the inequality holds for the expectation over all y, this means that the inequality holds for some fixed $y' \in \{0,1\}^n$. Let $S = \{i | y'_i = 1\}$.

$$\|\sum_{i\neq j} a_{ij}\sigma_i\sigma_j\|_p \leq 4\|\sum_{j\neq i} a_{ij}\sigma_i y'_i\sigma_j(1-y'_j)\|_p = 4\|\sum_{i\in S}\sum_{j\notin S} a_{ij}\sigma_i\sigma_j\|_p$$

But as the sets S and S^c are disjoint (S^c representing the complement of S), the two groups can be viewed as separate, i.e.

$$= 4\|\sum_{i\in S}\sum_{j\notin S} a_{ij}\sigma_i\sigma'_j\| = 4\|\mathop{\mathbb{E}}_{\sigma_S\sigma_{\bar{S}}}\sum_{i,j} a_{ij}\sigma_i\sigma'_j\|_p$$

where we added back the expectation of the missing terms, which is zero, and finally apply Jensen to obtain that the above is at most

$$4\|\sum_{i,j} a_{ij}\sigma_i\sigma'_j\|_p$$

□

4 Alon's Lower Bound

Theorem 2 (Alon [3])**.** *For every $N > 1$ there exists a set of points $x_0, \ldots, x_N \in \mathbb{R}^N$ such that any embedding into ℓ_2^m with distortion at most $1+\varepsilon$ for $\frac{1}{\sqrt{N}} < \varepsilon < \frac{1}{2}$ must have $m = \Omega(\frac{1}{\varepsilon^2 \log\frac{1}{\varepsilon}\log N})$.*

Proof. Let the points be $x_0 = 0$, and $x_i = e_i$ for $i > 0$, and let v_i be the image of e_i in the m-dimensional space. Now consider Π, whose columns are v_i, and notice the following.

- $|1 - \|v_i\|| < \varepsilon$
- $\|v_i - v_j\|^2 = \|v_i\|^2 + \|v_j\|^2 - 2 < v_i, v_j > \implies |\langle v_i, v_j\rangle| = O(\varepsilon)$

Thus, another way of thinking about Π is, once we rescale the v_i to be unit vector, is an ε-incoherent matrix, so this proof is also giving a lower-bound on the size of the ε-incoherent matrices.

Now note

$$\Pi^T\Pi = \begin{bmatrix} O(\varepsilon) & 0 & ... & O(\varepsilon) & O(\varepsilon) \\ O(\varepsilon) & 1 & ... & O(\varepsilon) & O(\varepsilon) \\ ... & ... & ... & ... & ... \\ O(\varepsilon) & O(\varepsilon) & ... & 1 & O(\varepsilon) \\ O(\varepsilon) & O(\varepsilon) & ... & O(\varepsilon) & 1 \end{bmatrix} \text{ (m blocks each n}\times\text{n)},$$

or in other words, $\Pi^T\Pi$ is an ε-near identity matrix, with rank at most m. We now use this lemma, also by Alon.

Lemma 3. *Any $n \times n$ ε-near identity with rank m must have $m \gtrsim \frac{1}{\varepsilon^2 \log \frac{1}{\varepsilon} \log n}$*

Clearly, the proof of Alon's lower bound follows immediately from the lemma. Proving this lemma requires another lemma.

Lemma 4. *If A, an ε-near identity is symmetric (note that $\Pi^T\Pi$ is symmetric, so this is all we need), then $m \geq \frac{n}{1+\varepsilon^2(n-1)}$*

Proof. Note that as the matrix is symmetric, it has m non-zero eigenvalues, $\lambda_1, ...\lambda_m$. We notice the following.

- $\sum_i \lambda_i^2 = \|A\|_F^2 \leq n + n(n-1)\varepsilon^2$
- $(\sum_i \lambda_i)^2 = (tr(A))^2 = n^2$
- (Cauchy-Schwarz) $(\sum_i \lambda_i)^2 \leq m \sum_i \lambda_i^2$
- $\implies n^2 \leq m(n + n(n-1)\varepsilon^2)$

Rearranging gives the desired inequality. □

We still do not have a good enough bound, however, so we bootstrap this lemma.

Bootstrapped Lemma 5. *Let $A = (a_{ij})$ be of rank m, and let $p : \mathbb{R} \mapsto \mathbb{R}$ a degree k polynomial, and we define $p(A) = (p(a_{ij}))$. Then, $rank(p(A)) \leq \binom{m+k}{k}$*

Proof. Suppose $v_1, ...v_m$ is a basis for the row space of A. If $p(z) = \sum_{i=0}^k \beta_i z^i$, then $p(a_{ij}) = \sum_{q=0}^k \beta_q (\sum_r \alpha_r v_{r,j})^q$, so from this expansion, one can realize that $(v_{1,t}^{d_1}, ...v_{m,t}^{d_t})$, where $1 \leq t \leq n$, and $\sum d_i \leq k$ spans the row space of $P(A)$. The total number of such vectors is $\binom{m+k}{k}$. This is a well-known combinatorial identity (the number of ways to place at most k balls into m bins). □

Given this lemma, now set $k = \frac{1}{2}\frac{\log n}{\log \frac{1}{\varepsilon}}$ (which is at least 1 for $\varepsilon > 1/\sqrt{n}$), $p(z) = z^k$. Now note that $p(\Pi^T\Pi)$ has off-diagonal entries equal to $\frac{1}{n}$. Putting the previous two lemmas together, we have the following.

$$\frac{n}{2} \leq rank(p(\Pi^T\Pi)) \leq \binom{m+k}{k}$$

Note we have $\binom{a}{b} \leq (e \cdot \frac{a}{b})^b$, so if we take the log of both sides, we obtain:

$$\log \frac{n}{2} \leq k \log \frac{e(m+1)}{k}$$

Rearranging gives the desired proof of the lemma, and hence the proof of Alon's lower bound, as desired. □

5 Improving the JL upper bound

Recall: T is a set of unit ℓ_2 vectors, a random sign matrix Π preserves all vectors in T simultaneously up to ε error, where $m \gtrsim \frac{\log T}{\varepsilon^2}$.

According to Gordon[5], who proved this result for random gaussian matrices, and Klartag and Mendelson[6], who proved it for random sign matrices (and other matrices with independent sub-gaussian entries), the bound can be improved to the following:

$$m \gtrsim \frac{g^2(T)+1}{\varepsilon^2}, g(T) = \mathbb{E} \sup_{x \in T} \langle g, x \rangle$$

where g is a random gaussian vector. This is a generalization of the JL lemma, since for any $x \in T$ it holds that $\langle g, x \rangle$ is a gaussian with unit variance. Thus

$$\mathbb{E} \sup_{x \in T} \langle g, x \rangle \stackrel{\text{def}}{=} \mathbb{E} \sup_{x \in T} g_x = \int_0^\infty \mathbb{P}(\sup_{x \in T} g_x > t) dt \leq \sqrt{\log |T|} + \sum_{x \in T} \int_{\sqrt{\log |T|}}^\infty \mathbb{P}(g_x > t) dt$$

is at most $O(\sqrt{\log |T|})$ (the last inequality was by the union bound), and thus $g^2(T) = O(\log |T|)$. However the bound can be much better than $O(\log |T|)$ depending on T, e.g. if the vectors in T fall into a small number of well-clustered sets (in which case the union bound is suboptimal).

References

[1] Gilles Pisier. Probabilistic methods in the geometry of Banach spaces. *Probability and Analysis*, Varenna (Italy) 1985. *Lecture Notes in Math.*, 1206:167–241, 1986.

[2] William B. Johnson and Joram Lindenstrauss. Extensions of Lipschitz mappings into a Hilbert space. *Contemporary Mathematics*, 26:189–206, 1984.

[3] Noga Alon. Problems and results in extremal combinatorics I. *Discrete Math.*, vol. 273, pp. 31–53, 2003.

[4] Victor Hugo de la Peña, Evarist Autor Gine. Decoupling: from dependence to independence. *Probability and its Applications* (New York). Springer-Verlag, New York, 1999.

[5] Yehoram Gordon. On Milman's inequality and random subspaces which escape through a mesh in $\mathbb{R}^n$. *Geom. Aspects of Funct. Anal.*, Israel seminar, Lecture Notes in Math., 1317, Springer-Verlag, 84–106, 1988.

[6] Bo'az Klartag and Shahar Mendelson. Empirical processes and random projections. *J. Funct. Anal.*, vol.225, pp. 229–245, 2005.