Lecture 5

<u>Review:</u> * From tensorization of variance to tensorization of ψ_X:

$$\mathrm{Var}\left(\sum_{i=1}^n X_i\right) = \sum_{i=1}^n \mathrm{Var}(X_i) \quad \Big| \quad \psi_{\sum_{i=1}^n X_i}(\lambda) = \sum_{i=1}^n \psi_{X_i}(\lambda)$$

* subGaussian tails:

$$\psi_X(\lambda) \le \frac{\lambda^2 v}{2} \longrightarrow \mathbb{P}\left(\sum_{i=1}^n X_i > \sqrt{2nv\log\frac{1}{\delta}}\right) \le \delta$$

* less than subGaussian tails:

(a) Poisson-tail (right)

$$\psi_X(\lambda) \le v(e^\lambda - 1 - \lambda)$$

(b) χ^2-tail (right)

$$\mathbb{P}\left(\sum_{i=1}^n X_i > \log\frac{1}{\delta} + \sqrt{cnv\log\frac{1}{\delta}}\right) \le \delta$$

(c) Exponential

$$\psi_X(\lambda) \le \frac{\lambda^2 v}{2}, \text{ for all } \lambda < \frac{c_1}{\sqrt{v}} \rightarrow \mathbb{P}\left(\sum_{i=1}^n X_i > \sqrt{cnv}\log\frac{1}{\delta}\right) \le \delta$$

* Concentration bounds from tensorization of $\psi_X(\lambda)$

I. Hoeffding inequality: $X_1, \dots, X_n$ indep, $\mathbb{E}[X_i] = 0$

Suppose $X_i \in [a_i, b_i]$. Then,

$$\mathbb{P}\left(\sum_{i=1}^n X_i > t\right) \le \exp\left(\frac{-2t^2}{\sum_{i=1}^n (b_i - a_i)^2}\right).$$

II. Bennett inequality

$$|X_i| < c, \quad \sigma^2 = \frac{1}{n} \cdot \sum_{i=1}^{n} \mathrm{Var}(X_i)$$

$$\mathbb{P}\left(\sum_{i=1}^{n} X_i > t \right) \le \exp\left(\frac{-t^2}{2\sigma^2 n + \frac{2ct}{3}} \right).$$

* Concentration using multiplicative property (martingale difference)

I. Azuma inequality

$X_1, \dots, X_n$ are zero-mean and mutually uncorrelated

$$|X_i| < c_i, \quad 1 \le i \le n$$

$$\mathbb{P}\left(\sum_{i=1}^{n} X_i > t \right) \le \exp\left(\frac{-t^2}{2\sum_{i=1}^{n} c_i^2} \right)$$

II. McDiarmid inequality

$$f: \mathcal{X}_1 \times \dots \times \mathcal{X}_n \to \mathbb{R}$$

$X_1, \dots, X_n$ are indep, denote $X = (X_1, \dots, X_n)$

Assume: f satisfies BDP with $\underline{c} = (c_1, \dots, c_n)$, i.e.,

$$f(x) - f(x') \le \sum_{i=1}^{n} c_i \mathbb{1}_{\{x_i \neq x_i'\}}$$

$$\mathbb{P}\left(f(X) - \mathbb{E} f(X) > t \right) \le \exp\left(\frac{-t^2}{2\sum_{i=1}^{n} c_i^2} \right)$$

Agenda: Applications

[A] A bound for $\mathbb{E}[\max_i X_i]$ for subGaussian X_i

[B] - longest increasing subsequence
- longest common subsequence
- chromatic number of Erdős-Rényi random graphs

[C] Balls and bins

[D] Concentration empirical measures
- introduction to chaining
- VC theory
- Glivenko-Cantelli theorem

$\Big\}$ next lecture

[A] A maximal inequality

Suppose $X_1, \dots, X_n$ are subGaussian with variance parameter σ^2. (no independence assumption!)

By Jensen's inequality, for $\lambda > 0$,

$$\lambda \mathbb{E}\left[\max_i X_i\right] \leq \log \mathbb{E}\left[e^{\lambda \max_i X_i}\right].$$

Further,

$$\mathbb{E}\left[e^{\lambda \max_i X_i}\right] = \mathbb{E}\left[\max_i e^{\lambda X_i}\right] \leq \mathbb{E}\left[\sum_{i=1}^n e^{\lambda X_i}\right]$$
$$= \sum_{i=1}^n \mathbb{E}\left[e^{\lambda X_i}\right].$$

Thus,

$$\lambda \mathbb{E}\left[\max_i X_i\right] \leq \log n \cdot e^{\lambda^2\sigma^2/2}$$
$$= \log n + \frac{\lambda^2\sigma^2}{2},$$

i.e.,

$$\boxed{\mathbb{E}\left[\max_i X_i\right] \leq \frac{\log n}{\lambda} + \frac{\lambda\sigma^2}{2} \leq \sigma\sqrt{\frac{\log n}{2}}}$$

$\boxed{B}$ Concentration of some interesting quantities

Applications of McDiarmid

(a) Longest Common Sequence

Let $X_1, \ldots, X_n, Y_1, \ldots, Y_n$ be iid P_{XY}. $(X_i, Y_i \in \mathcal{X})$

$$L = \max\{k \mid \exists\ i_1 < i_2 \ldots < i_k \text{ and } j_1 < j_2 < \ldots < j_k \text{ s.t. } X_{i_\ell} = Y_{j_\ell},\ 1 \le \ell \le k\}$$

Claim: $\mathbb{P}\left(|L - \mathbb{E}L| > t\right) \le 2e^{-t^2/8n}$

Pf. By changing any one pair (X_i, Y_i), L can change by at most 2. Use Hoeffding inequality.

Remark. $\mathbb{E}L = \frac{n}{|\mathcal{X}|^2}$ (check: Chvatal-Sankoff)

(b) Longest increasing sequence

$X_1, \ldots, X_n$ are iid P_X

$$f(X^n) = \max\{k \mid \exists\ i_1 < i_2 \ldots < i_k,\ X_{i_1} < X_{i_2} \ldots < X_{i_k}\}$$

f satisfies BDP with $(1, \ldots, 1)$.

(c) Erdös-Rényi random graphs

$G(n,p)$: Graph on n vertices, $(i,j) \in \mathcal{E}$ with prob. p - indep. for all pairs (i,j)

$$X_{ij} = \mathbb{1}(\exists \text{ edge b/w } i \text{ and } j), \quad 1 \le i < j \le n$$

A coloring of G is an assignment of colors to the vertices of G s.t. no two adjacent vertices have the same color.

The chromatic number of G is the min. # of colors required for coloring G.

$\chi(G(n,p)) \equiv$ chromatic # of $G(n,p)$

<u>Claim:</u> $\mathbb{P}\left(\left|\chi(G(n,p)) - \mathbb{E}\,\chi(G(n,p))\right| > t\right) \le 2\, e^{-\frac{t^2}{2n}}$.

<u>Remark:</u> $\mathbb{E}\,\chi(G(n,p)) \approx \begin{cases} \frac{n}{\log n} & \text{when } p > \frac{\log n}{n} \\ c & \text{when } p < \frac{\log n}{n} \end{cases}$

$\boxed{C}$ <u>Balls and bins</u>

* Throw <u>m balls</u> independently and uniformly at random into <u>n bins</u>

$$Z = \# \text{ of empty bins} = \sum_{i=1}^{n} \underbrace{\mathbb{1}(\text{bin } i \text{ is empty})}_{X_i}$$

$$\mathbb{E}Z = n\,\mathbb{P}(X_i = 1) = n\left(1 - \frac{1}{n}\right)^m \le e^{-\frac{m}{n} + \log n}$$

Thus, $\mathbb{E}Z \ge 1$ roughly only when $m \ll n \log n$.

Note that $X_1, \dots, X_n$ are not indep.

Let B_j = index of the bin where ball j falls $\in \{1, \dots, n\}$, for $1 \le j \le m$.

Then, $Z = f(B_1, \dots, B_m)$. Clearly, f satisfies BDP with $(1, \dots, 1)$. Thus, by McDiarmid's inequality,

$$\mathbb{P}(Z - \mathbb{E}Z > t) \le \exp\left(-\frac{t^2}{2m}\right).$$

that is

$$\mathbb{P}\left(Z > \mathbb{E}Z + \sqrt{2m\log\frac{1}{\delta}}\right) \leq \delta$$

$$\Leftrightarrow \mathbb{P}\left(Z \leq \mathbb{E}Z + \sqrt{2m\log\frac{1}{\delta}}\right) \geq 1-\delta.$$

To study the max. m s.t. no bin is empty, it is better to consider

$$Z' = \#\text{ of non-empty bins}$$

and bound $\mathbb{P}(Z' \geq n)$. As before

$$\mathbb{P}\left(Z' - \mathbb{E}Z' > \sqrt{2m\log\frac{1}{\delta}}\right) \leq \delta$$

Also

$$\mathbb{E}Z' = n\left(1 - \left(1-\frac{1}{n}\right)^m\right) \approx m$$

Thus, $\mathbb{P}(Z' \geq n)$ is small if (roughly)

$$m + \sqrt{m} \leq n. \quad \left(\begin{array}{l} m+\sqrt{m} \leq (\sqrt{m}+0.5)^2 \leq n \\ \Leftrightarrow \sqrt{m} \leq \sqrt{n} - 0.5 \\ \Leftarrow \sqrt{m} \leq \frac{\sqrt{n}}{2} \end{array}\right)$$

$$\Uparrow$$

$$m \leq \frac{n}{c}.$$

That is, up to $m = \frac{n}{c}$ there is at least one nonempty bin.

Also, similarly $\mathbb{P}\left(Z' < \mathbb{E}Z - \sqrt{2m\log\frac{1}{\delta}}\right) \leq \delta$.

Thus, if $m - \sqrt{m} > n$, $\mathbb{P}(Z' < n) \leq \delta$. This again holds once $m \geq c'n$.

* <u>What happens for $m = n$?</u>

We seek to study the max. load in this regime.

$$X_i^{(m)} \overset{\text{def}}{=} \#\text{ of balls in bin } i$$

$$X_i^{(m)} \sim \text{Bin}\left(m, \frac{1}{n}\right)$$

Thus, $$\mathbb{P}\left(\max_i X_i^{(m)} > t \right) \leq n\, \mathbb{P}\left(X_1^{(m)} > t\right)$$
(by union bound)

By Bennett inequality,

$$\mathbb{P}\left(X_1^{(m)} - \frac{m}{n} > t\right) \leq \exp\left(-\frac{t^2}{2\frac{m}{n} + \frac{2t}{3}}\right)$$

In particular, for $m=n$,

$$\mathbb{P}\left(X_1^{(n)} - 1 > c \log \frac{n}{\delta}\right) \leq \frac{\delta}{n}$$

$$\Rightarrow \mathbb{P}\left(\max_{1\leq i\leq n} X_i^{(m)} > 1 + c \log \frac{n}{\delta}\right) \leq \delta$$

$$\text{Max. load} = O(\log n)$$

<u>Reading assignment</u>: Poisson approximation
(Mitchenmacher and Upfal, "Probability and computing")