Lecture 3

Last time

- Chernoff Method
- Sub-Gaussians
- Hoeffding's inequality

Today

- Robust mean estimation
- Chernoff's inequality
- Degree of random graphs.

Robust Mean Estimation

Using Hoeffding's inequality we can obtain confidence intervals for mean estimation of sub-Gaussians

If $X_1, \dots, X_n$ iid σ^2-sub-Gaussians, then

$\hat{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$ satisfies that $\forall p > 0$

$$\mathbb{P}\left(|\hat{x} - \mathbb{E}X_1| \le \sqrt{2\sigma^2 \log(1/p)} \right) \ge 1 - p.$$

A natural question is whether one can derive a similar guarantee for some estimator $\hat{x}$ if the x_i are not sub-Gaussian?

The answer is yes! Try to think how before reading the following.

The intuition is simple: one can bootstrap an estimator by ignoring outliers

Theorem: Consider $X_1, \dots, X_n$ iid with $\mathbb{E}X_1 = \mu$ and $\mathrm{Var}(X) = \sigma^2$. Fix $p \in [0,1]$. Split the samples into $k = 18\log(1/p)$ (equally sized) bins and let the empirical means be $\hat{X}_j$ for $j = 1, \dots, k$. Then, $\hat{x} = \mathrm{median}(\hat{X}_1, \dots, \hat{X}_k)$ satisfies

$$\mathbb{P}\left(|\hat{x} - \mu| \le \sqrt{\frac{3\sigma^2 \log(1/p)}{n}}\right) \ge 1 - p. \quad \dashv$$

Recall that $\#\{i \mid \hat{X}_i \le \mathrm{median}\} \ge \frac{1}{2}$,
$\#\{i \mid \hat{X}_i \ge \mathrm{median}\} \ge \frac{1}{2}$.

Remarks:

1. We only require second moments and so the MGF might not exist.
2. The estimator depends on p. This is unavoidable (see Exercise 2.17 of Vershynin's).

Proof: First the main idea: suppose that we have a $t > 0$ s.t.

Then, $|\text{median}(\hat{X}_1, \dots, \hat{X}_k) - \mu| \le t$. (Why?)
Consider the events

$$\mathcal{E}_i = \left\{ |\hat{X}_i - \mu| \le \sqrt{\frac{3\sigma^2 k}{n}} \right\}.$$

If more than half of these events hold, then, we get $|\hat{X} - \mu| \le \sqrt{\frac{3\sigma^2 k}{n}}$. This equivalent to

$$\frac{1}{k} \sum \mathbb{1}_{\mathcal{E}_i} > \frac{1}{2} \qquad (\star)$$

($\mathbb{1}_{\mathcal{E}_i}$: indicator of $\mathcal{E}_i$)

By Chebychev

$$\mathbb{P}(\mathcal{E}_i^c) \le \frac{\text{Var}(X_i)}{\frac{3\sigma^2 k}{n}} = \frac{\frac{\sigma^2}{(n/k)}}{\frac{3\sigma^2 k}{n}} = \frac{1}{3}.$$

So $\mathbb{E}\,\mathbb{1}_{\mathcal{E}_i} > \frac{2}{3}$, and if $(\star)$ does not hold

$$\frac{1}{k} \sum \mathbb{1}_{\mathcal{E}_i} \le \frac{1}{2} \Rightarrow \frac{1}{k} \sum \mathbb{1}_{\mathcal{E}_i} \le \frac{2}{3} - \frac{1}{6}$$

$$\le \mathbb{E}\,\mathbb{1}_{\mathcal{E}_i} - \frac{1}{6}$$

$$\Rightarrow \left| \frac{1}{k} \sum \mathbb{1}_{\mathcal{E}_i} - \mathbb{E}\,\mathbb{1}_{\mathcal{E}_i} \right| \ge \frac{1}{6}$$

Thus,

$$\mathbb{P}\left[\frac{1}{k} \sum_{i=1}^{k} \mathbb{1}_{\mathcal{E}_i} \le \frac{1}{2} \right] \le \mathbb{P}\left[\left| \frac{1}{k} \sum \mathbb{1}_{\mathcal{E}_i} - \mathbb{E}\,\mathbb{1}_{\mathcal{E}_i} \right| \ge \frac{1}{6} \right]$$

Hoeffding → $\leq \exp\left(-\frac{k}{18}\right)$. □

One of the potential final projects has to do with how to do this in higher dimensions than one.

Chernoff's inequality.

A weakness of Hoeffding's for bounded r.v. is that it only takes into account the worst possible spread $(b-a)$ instead of the average one.

Example: If $X_i \sim Ber(p) = \begin{cases} 1 & \text{w.p. } p \\ 0 & \text{w.p. } 1-p \end{cases}$

Then, as $p \to 0$, concentration is better, yet Hoeffding's does not improve. →

A much better bound holds for Bernoullis.

Theorem (Chernoff's inequality)

Let $X_i \sim Ber(p_i)$ be iid. Then, $S_n = \sum_{i=1}^{n} X_i$ has mean $ES_n = \Sigma p_i = \mu$ and

$$\mathbb{P}(S_n \geq t) \leq e^{-\mu}\left(\frac{e\mu}{t}\right)^t \quad \forall t \geq \mu. \quad \dagger$$

Proof: We apply the same idea as in the proof of Hoeffding's:

$$\mathbb{P}(S_n \geq t) = \mathbb{P}(e^{\lambda S_n} \geq e^{\lambda t})$$

Markov's → $\leq e^{-\lambda t}\,\mathbb{E}\,e^{\lambda S_n}$

Independence → $= e^{-\lambda t}\prod_{i=1}^n \mathbb{E}\,e^{\lambda X_i}$

$$= e^{-\lambda t}\prod\left(p_i e^{\lambda} + (1-p_i)\right)$$

To combine the terms in the product we use $1+x \leq e^x$, which gives

$$\mathbb{P}(S_n \leq t) \leq e^{-\lambda t}\prod \exp\left((e^{\lambda}-1)p_i\right)$$
$$= e^{-\lambda t}\exp\left((e^{\lambda}-1)\sum_{i=1}^n p_i\right)$$
$$= \exp\left(-\lambda t + (e^{\lambda}-1)\mu\right)$$

Minimizing for λ yields $\lambda = \ln(t/\mu)$ (check!) and so

(By assumption $t \geq \mu$)

$$\mathbb{P}(S_n \leq t) \leq \exp\left(-\ln(t/\mu)\,t + t - \mu\right)$$
$$= e^{-\mu}\left(\frac{e\mu}{t}\right)^t$$

Remarks:

1. One can extend the result

$$\mathbb{P}(S_n \le t) \le e^{-\mu}\left(\frac{e\mu}{t}\right)^t \quad \forall\, 0 \le t \le \mu.$$

2. Chernoff is optimal in that

$$\mathbb{P}(S_N \ge t) \ge \left(\frac{\mu}{t}\right)^t \quad \forall\, 1 \le t \le n.$$

3. When t is large (Large deviations)

$$e^{-\mu}\left(\frac{e\mu}{t}\right)^t \sim t^{-t} \le e^{-t\log t}$$

which is much heavier that $e^{-t^2/2}$.

4. When $t \sim \mu$ (small deviations)

say $t = (1+\delta)\mu$ (δ small)

$$e^{-\mu}\left(\frac{e\mu}{t}\right)^t = e^{-\mu}\left(\frac{e}{(1+\delta)}\right)^{(1+\delta)\mu}$$

$$= e^{\delta\mu}\left(\frac{1}{1+\delta}\right)^{(1+\delta)\mu}$$

$$= \exp\left[\mu\left(\delta - (1+\delta)\log(1+\delta)\right)\right]$$

$$\le \exp\left(-\frac{\delta^2}{6}\mu\right)$$

$\delta + \frac{\delta^2}{2} + o(\delta^3)$ (Taylor expansion)