Theoretical Statistics. Lecture 4.

Peter Bartlett

1. Concentration inequalities.

Outline of today's lecture

We have been looking at **deviation inequalities**, i.e., bounds on tail probabilities like $P(X_n \geq t)$ for some statistic X_n.

1. Using moment generating function bounds, for sums of independent r.v.s:
   Chernoff; Hoeffding; sub-Gaussian, sub-exponential random variables; Bernstein.

   Today: Johnson-Lindenstrauss.

2. Martingale methods:
   Hoeffding-Azuma, bounded differences.

Review. Chernoff technique

Theorem: For $t > 0$:

$$P(X - \mathbf{E}X \geq t) \leq \inf_{\lambda > 0} e^{-\lambda t} M_{X-\mu}(\lambda).$$

Theorem: [Hoeffding's Inequality] For a random variable $X \in [a, b]$ with $\mathbf{E}X = \mu$ and $\lambda \in \mathbb{R}$,

$$\ln M_{X-\mu}(\lambda) \leq \frac{\lambda^2 (b-a)^2}{8}.$$

Review. Sub-Gaussian, Sub-Exponential Random Variables

Definition: X is **sub-Gaussian** with parameter σ^2 if, for all $\lambda \in \mathbb{R}$,

$$\ln M_{X-\mu}(\lambda) \leq \frac{\lambda^2 \sigma^2}{2}.$$

Definition: X is **sub-exponential** with parameters (σ^2, b) if, for all $|\lambda| < 1/b$,

$$\ln M_{X-\mu}(\lambda) \leq \frac{\lambda^2 \sigma^2}{2}.$$

Review. Sub-Exponential Random Variables

Theorem: For X sub-exponential with parameters (σ^2, b),

$$P(X \geq \mu + t) \leq \begin{cases} \exp\left(-\frac{t^2}{2\sigma^2}\right) & \text{if } 0 \leq t \leq \sigma^2/b, \\ \exp\left(-\frac{t}{2b}\right) & \text{if } t > \sigma^2/b. \end{cases}$$

- For independent X_i, sub-exponential with parameters (σ_i^2, b_i), the sum $X = X_1 + \cdots + X_n$ is sub-exponential with parameters $\left(\sum_i \sigma_i^2, \max_i b_i\right)$.
- Example: $X \sim \chi_1^2$ is sub-exponential with parameters $(4, 4)$.

Sub-Exponential Random Variables: Example

Theorem: [Johnson-Lindenstrauss] For m points $x_1, \ldots, x_m$ from $\mathbb{R}^d$, there is a projection $F : \mathbb{R}^d \to \mathbb{R}^n$ that preserves distances in the sense that, for all x_i, x_j,

$$(1-\delta)\|x_i - x_j\|_2^2 \leq \|F(x_i) - F(x_j)\|_2^2 \leq (1+\delta)\|x_i - x_j\|_2^2,$$

provided that $n > (16/\delta^2)\log m$.

That is, we can embed these points in $\mathbb{R}^n$ and approximately maintain their distance relationships, provided that n is not too small. Notice that n is independent of the ambient dimension d, and depends only logarithmically on the number of points m.

Johnson-Lindenstrauss

Applications: dimension reduction to simplify computation (nearest neighbor, clustering, image processing, text processing).
Analysis of machine learning methods: separable by a large margin in high dimensions implies it's really a low-dimensional problem after all.

Johnson-Lindenstrauss Embedding: Proof

We use a random projection:

$$F(x) = \frac{1}{\sqrt{n}} Yx,$$

where $Y \in \mathbb{R}^{n \times d}$ has independent $N(0,1)$ entries.

Let Y_i denote the ith row, for $1 \le i \le n$. It has a $N(0, I)$ distribution, so $Y_i^T x / \|x\|_2 \sim N(0,1)$. Thus,

$$Z = \frac{\|Yx\|_2^2}{\|x\|_2^2} = \sum_{i=1}^{n} \left(Y_i^T x / \|x\| \right)^2 \sim \chi_n^2.$$

Johnson-Lindenstrauss Embedding: Proof

Since $Z \sim \chi_n^2$ is the sum of n independent sub-exponential $(4, 4)$ random variables, it is sub-exponential $(4n, 4)$. And we have that for $0 < t < n$,

$$P(|Z - 1| \geq t) \leq 2\exp(-t^2/(8n)).$$

Hence, for $0 < \delta < 1$,

$$P\left(\left|\frac{\|Yx\|_2^2}{n\|x\|_2^2} - 1\right| \geq \delta\right) \leq 2\exp(-n\delta^2/8)$$

$$\Leftrightarrow P\left(\frac{\|F(x)\|_2^2}{\|x\|_2^2} \notin [1-\delta, 1+\delta]\right) \leq 2\exp(-n\delta^2/8).$$

Johnson-Lindenstrauss Embedding: Proof

Applying this to the $\binom{m}{2}$ distinct pairs $x = x_i - x_j$, and using the union bound gives

$$P\left(\exists i \neq j \text{ s.t. } \frac{\|F(x_i - x_j)\|_2^2}{\|x_i - x_j\|_2^2} \notin [1-\delta, 1+\delta]\right) \leq 2\binom{m}{2} \exp(-n\delta^2/8).$$

Thus, for $n > 16/\delta^2 \log(m)$, this probability is strictly less than 1, so there exists a suitable mapping.

In fact, we can choose a random projection in this way and ensure that the probability that it does not satisfy the approximate isometry property is no more than ϵ for $n > 16/\delta^2 \log(m/\epsilon)$.

Concentration Bounds for Martingale Difference Sequences

Next, we're going to consider concentration of martingale difference sequences. The application is to understand how tails of $f(X_1, \ldots, X_n) - \mathbf{E}f(X_1, \ldots, X_n)$ behave, for some function f.

[e.g., in the homework, we have that f is some measure of the performance of a kernel density estimator.] If we write

$$f(X_1, \ldots, X_n) - \mathbf{E}f(X_1, \ldots, X_n)$$

$$= \sum_{i=1}^{n} \mathbf{E}[f(X_1, \ldots, X_n)|X_1, \ldots, X_i] - \mathbf{E}[f(X_1, \ldots, X_n)|X_1, \ldots, X_{i-1}],$$

then we have represented this deviation as a *martingale difference sequence*.

Martingales

Definition: A sequence Y_n of random variables adapted to a filtration $\mathcal{F}_n$ is a **martingale** if, for all n,

$$\mathbf{E}|Y_n| < \infty$$
$$\mathbf{E}[Y_{n+1}|\mathcal{F}_n] = Y_n.$$

$\mathcal{F}_n$ is a **filtration** means these σ-fields are nested: $\mathcal{F}_n \subseteq \mathcal{F}_{n+1}$.

Y_n is **adapted to** $\mathcal{F}_n$ means that each Y_n is measurable with respect to $\mathcal{F}_n$.

e.g. $\mathcal{F}_n = \sigma(Y_1, \ldots, Y_n)$, the σ-field generated by the first n variables. Then we say Y_n is a martingale sequence.

e.g. $\mathcal{F}_n = \sigma(X_1, \ldots, X_n)$. Then Y_n is a martingale sequence wrt X_n.

Martingale Difference Sequences

Definition: A sequence D_n of random variables adapted to a filtration $\mathcal{F}_n$ is a **martingale difference sequence** if, for all n,

$$\mathbf{E}|D_n| < \infty$$
$$\mathbf{E}[D_{n+1}|\mathcal{F}_n] = 0.$$

e.g., $D_n = Y_n - Y_{n-1}$.

$$\begin{aligned}\mathbf{E}[D_{n+1}|\mathcal{F}_n] &= \mathbf{E}[Y_{n+1}|\mathcal{F}_n] - \mathbf{E}[Y_n|\mathcal{F}_n] \\ &= \mathbf{E}[Y_{n+1}|\mathcal{F}_n] - Y_n = 0\end{aligned}$$

(because Y_n is measurable wrt $\mathcal{F}_n$, and because of the martingale property).
Hence, $Y_n - Y_0 = \sum_{i=1}^n D_i$.

Martingale Difference Sequences: the Doob construction

$$\text{Define} \qquad \begin{aligned} X &= (X_1, \ldots, X_n), \\ X_1^i &= (X_1, \ldots, X_i), \\ Y_0 &= \mathbf{E} f(X), \\ Y_i &= \mathbf{E}[f(X)|X_1^i]. \end{aligned}$$

$$\text{Then} \qquad f(X) - \mathbf{E} f(X) = Y_n - Y_0 = \sum_{i=1}^{n} D_i,$$

where $D_i = Y_i - Y_{i-1}$. Also, Y_i is a martingale w.r.t. X_i, and hence D_i is a martingale difference sequence. Indeed (because $\mathbf{E}X = \mathbf{E}\mathbf{E}[X|Y]$),

$$\mathbf{E}[Y_{i+1}|X_1^i] = \mathbf{E}\left[\mathbf{E}[f(X)|X_1^{i+1}]\middle| X_1^i\right] = \mathbf{E}[f(X)|X_1^i] = Y_i.$$

Martingale Difference Sequences: another example

[An aside:] Consider two densities f and g, with g absolutely continuous w.r.t. f. Suppose X_n are drawn i.i.d. from f, and Y_n is the likelihood ratio,

$$Y_n = \prod_{i=1}^{n} \frac{g(X_i)}{f(X_i)}.$$

Then Y_n is a martingale w.r.t. X_n. Indeed,

$$\begin{aligned}\mathbf{E}[Y_{n+1}|X_1^n] = \mathbf{E}\left[\prod_{i=1}^{n+1} \frac{g(X_i)}{f(X_i)}\right| \left. X_1^n\right] &= \mathbf{E}\left[\frac{g(X_{n+1})}{f(X_{n+1})}\right] \prod_{i=1}^{n} \frac{g(X_i)}{f(X_i)} \\ &= \prod_{i=1}^{n} \frac{g(X_i)}{f(X_i)} = Y_n,\end{aligned}$$

because $\mathbf{E}[g(X_{n+1})/f(X_{n+1})] = 1$.

Concentration Bounds for Martingale Difference Sequences

Theorem: Consider a martingale difference sequence D_n (adapted to a filtration $\mathcal{F}_n$) that satisfies

$$\text{for } |\lambda| \leq 1/b_n \text{ a.s., } \mathbf{E}\left[\exp(\lambda D_n) \middle| \mathcal{F}_{n-1}\right] \leq \exp(\lambda^2 \sigma_n^2 / 2).$$

Then $\sum_{i=1}^n D_i$ is sub-exponential, with $(\sigma^2, b) = (\sum_{i=1}^n \sigma_i^2, \max_i b_i)$.

$$P\left(\left|\sum_i D_i\right| \geq t\right) \leq \begin{cases} 2\exp(-t^2/(2\sigma^2)) & \text{if } 0 \leq t \leq \sigma^2/b \\ 2\exp(-t/(2b)) & \text{if } t > \sigma^2/b. \end{cases}$$

Concentration Bounds for Martingale Difference Sequences

Proof:

$$\mathbf{E} \exp\left(\lambda \sum_i D_i\right) = \mathbf{E}\left[\exp\left(\lambda \sum_{i=1}^{n-1} D_i\right) \mathbf{E}\left[\exp(\lambda D_n) | \mathcal{F}_{n-1}\right]\right]$$
$$\leq \mathbf{E}\left[\exp\left(\lambda \sum_{i=1}^{n-1} D_i\right)\right] \exp(\lambda^2 \sigma_n^2/2),$$

provided $|\lambda| < b$. Iterating shows that $\sum_i D_i$ is sub-exponential.

Concentration Bounds for Martingale Difference Sequences

Theorem: Consider a martingale difference sequence D_i with $|D_i| \leq B_i$ a.s. Then

$$P\left(\left|\sum_i D_i\right| \geq t\right) \leq 2\exp\left(-\frac{2t^2}{\sum_i B_i^2}\right).$$

Proof:

It suffices to show that

$$\mathbf{E}\left[\exp(\lambda D_i) | \mathcal{F}_{i-1}\right] \leq \exp(\lambda^2 B_i^2/2)$$

But $|D_i| \leq B_i$ a.s., so the conditioned variable $(D_i|\mathcal{F}_{i-1}) \leq B_i$ a.s., so it is sub-Gaussian with parameter $\sigma_i^2 = B_i^2$.

Bounded Differences Inequality

Theorem: Suppose $f : \mathcal{X}^n \to \mathbb{R}$ satisfies the following **bounded differences inequality**:

for all $x_1, \ldots, x_n, x_i' \in \mathcal{X}$,

$$|f(x_1, \ldots, x_n) - f(x_1, \ldots, x_{i-1}, x_i', x_{i+1}, \ldots, x_n)| \leq B_i.$$

Then

$$P\left(|f(X) - \mathbf{E}f(X)| \geq t\right) \leq 2\exp\left(-\frac{2t^2}{\sum_i B_i^2}\right).$$

Bounded Differences Inequality

Proof: Use the Doob construction.

$$Y_i = \mathbf{E}[f(X)|X_1^i],$$
$$D_i = Y_i - Y_{i-1},$$
$$f(X) - \mathbf{E}f(X) = \sum_{i=1}^{n} D_i.$$

Then

$$\begin{aligned}|D_i| = |Y_i - Y_{i-1}| &= \left|\mathbf{E}[f(X)|X_1^i] - \mathbf{E}[f(X)|X_1^{i-1}]\right| \\ &= \left|\mathbf{E}\left[\mathbf{E}[f(X)|X_1^i] - f(X)\middle| X_1^{i-1}\right]\right| \leq B_i.\end{aligned}$$

Examples: Rademacher Averages

For a set $A \subset \mathbb{R}^n$, consider

$$Z = \sup_{a \in A} \langle \epsilon, a \rangle,$$

where $\epsilon = (\epsilon_1, \ldots \epsilon_n)$ is a sequence of i.i.d. uniform $\{\pm 1\}$ random variables. Define the **Rademacher complexity** of A as $R(A) = \mathbf{E}Z$. [This is a measure of the size of A.] The bounded differences approach implies that Z is concentrated around $R(A)$:

Theorem: Z is sub-Gaussian with parameter $4 \sum_i \sup_{a \in A} a_i^2$.

Proof:

Write $Z = f(\epsilon_1, \ldots, \epsilon_n)$, and notice that a change of ϵ_i can lead to a change in Z of no more than $B_n = \sup_{a \in A} 2|a_i|$. The result follows.

Examples: Empirical Processes

For a class F of functions $f : \mathcal{X} \to [0,1]$, suppose that $X_1, \ldots, X_n, X$ are i.i.d. on $\mathcal{X}$, and consider

$$Z = \sup_{f \in F} \left| \mathbf{E} f(X) - \frac{1}{n} \sum_{i=1}^{n} f(X_i) \right| = \Bigg\| \underbrace{Pf - P_n f}_{\text{emp proc}} \Bigg\|_F .$$

If Z converges to 0, this is called a *uniform law of large numbers*. Here, we show that Z is concentrated about $\mathbf{E}Z$:

Theorem: Z is sub-Gaussian with parameter $1/n$.

Proof:

Write $Z = g(X_1, \ldots, X_n)$, and notice that a change of X_i can lead to a change in Z of no more than $B_n = 1/n$. The result follows.