EECS281B/STAT241B: Advanced Topics in Statistical Learning Spring 2009

Lecture 13 — March 4

Lecturer: Martin Wainwright *Scribe: Gia Vinh Anh Pham*

Note: These lecture notes are still rough, and have only have been mildly proofread.

This is the danger environment.

13.1 Recap

Theorem 13.1 (Hoeffding). *Say $X_1, \ldots, X_n$ are independent random variables with $a_i \leq X_i \leq b_i$*

$$\mathbb{P}\left[\left|\frac{1}{n}\sum_{i=1}^{n}(X_i - \mathbb{E}[X_i])\right| > \epsilon\right] \leq 2\exp\left\{\frac{-2n\epsilon^2}{\frac{1}{n}\sum_{i=1}^{n}(b_i - a_i)^2}\right\}$$

Proof: Using Chernoff inequality,

$$\begin{aligned}
\mathbb{P}\left[\frac{1}{n}\sum_{i=1}^{n}(X_i - \mathbb{E}[X_i]) > \epsilon\right] &\leq \exp(-\lambda n\epsilon)\mathbb{E}\left[\exp\left(\lambda\sum_{i=1}^{n}(X_i - \mathbb{E}[X_i])\right)\right] \\
&= \exp(-\lambda n\epsilon)\mathbb{E}\left[\prod_{i=1}^{n}\exp\left(\lambda(X_i - \mathbb{E}[X_i])\right)\right] \\
&= \exp(-\lambda n\epsilon)\prod_{i=1}^{n}\mathbb{E}\left[\exp\left(\lambda(X_i - \mathbb{E}[X_i])\right)\right]
\end{aligned}$$

Since $e^{\lambda t}$ is a convex function, and $\mathbb{E}\left[X_i - \mathbb{E}[X_i]\right] = 0$ we have

$$\mathbb{E}\left[\exp\left(\lambda(X_i - \mathbb{E}[X_i])\right)\right] \leq \frac{-a_i}{b_i - a_i}e^{b_i\lambda} + \frac{b_i}{b_i - a_i}e^{a_i\lambda} = e^{a_i\lambda}\left(1 + \frac{a_i}{b_i - a_i} - \frac{a_i}{b_i - a_i}e^{b_i\lambda}\right)$$

Define $g(u) = \log(1 - c + ce^u) - cu$ where $c = \frac{-a_i}{b_i - a_i}$, then the above inequality is equivalent to

$$\mathbb{E}\left[\exp\left(\lambda(X_i - \mathbb{E}[X_i])\right)\right] \leq e^{g(\lambda(a_i - b_i)}$$

We note that

$$\begin{aligned}
g(0) &= 0 \\
g'(u) &= \frac{ce^u}{1 - c + ce^u} - c \Rightarrow g'(0) = 0 \\
g''(u) &= \frac{ce^u(1-c)}{(1 - c + ce^u)^2} \leq \frac{\frac{1}{4}(ce^u + 1 - c)^2}{(1 - c + ce^u)^2} = \frac{1}{4}, \forall u
\end{aligned}$$

Using Taylor series expansion we can write

$$g(u) = g(0) + g'(0)u + \frac{g''(\bar{u})}{2}u^2 \leq \frac{1}{8}u^2$$

So,

$$\mathbb{E}\left[\exp\left(\lambda(X_i - \mathbb{E}[X_i])\right)\right] \leq e^{g(\lambda(a_i - b_i)} \leq \frac{1}{8}\lambda^2(a_i - b_i)^2$$

Therefore

$$\mathbb{P}\left[\frac{1}{n}\sum_{i=1}^{n}(X_i - \mathbb{E}[X_i]) > \epsilon\right] \leq \exp\left(-\lambda n\epsilon + \frac{\lambda^2}{8}\sum_{i=1}^{n}(b_i - a_i)^2\right) \leq \exp\left(\frac{-2n\epsilon^2}{\frac{1}{n}\sum_{i=1}^{n}(b_i - a_i)^2}\right)$$

□

13.2 Martingale inequalities

(Way of relaxing independence assumptions)

Definition: A sequence of integrable random variables $Z_1, Z_2, \ldots$ is a martingale if

$$\mathbb{E}\left[Z_{n+1}|Z_1, \ldots, Z_n\right] = Z_n, \forall n = 1, 2, \ldots$$

Example:
Say $V_1, V_2, \ldots$ are independent zero mean random variables. Define $S_n = \sum_{i=1}^{n} V_i$. This is a martingale:

$$\begin{aligned}\mathbb{E}\left[S_{n+1}|S_1, \ldots, S_n\right] &= \mathbb{E}\left[S_n + V_{n+1}|S_1, \ldots, S_n\right] \\ &= S_n + \mathbb{E}\left[V_{n+1}|S_1, \ldots, S_n\right] \\ &= S_n + \mathbb{E}\left[V_{n+1}\right] \text{ (by independence)} \\ &= S_n\end{aligned}$$

Definition: A sequence of random variables $V_1, V_2, \ldots$ is a martingale difference sequence (MDS) if

$$\mathbb{E}\left[V_{n+1}|V_1, \ldots, V_n\right] = 0, \forall n = 1, 2, \ldots$$

Proposition: If $Z_1, Z_2, \ldots$ is a martingale then $\{V_n | V_{n+1} = Z_{n+1} - Z_n\}$ is a MDS.

Lemma 13.2 (Azuma-Hoeffding). *Say $V_1, V_2, \ldots$ is a martingale difference sequence with $a_i \leq V_i \leq b_i, i = 1, \ldots, n$. Then $\forall \epsilon > 0$*

$$\mathbb{P}\left[\left|\frac{1}{n}\sum_{i=1}^{n} V_i\right| > \epsilon\right] \leq 2\exp\left\{\frac{-2n\epsilon^2}{\frac{1}{n}\sum_{i=1}^{n}(b_i - a_i)^2}\right\}$$

Proof: Claim that $\forall \lambda \in R, i = 1, 2 \ldots$:

$$\mathbb{E}\left[\exp(\lambda V_i)|V_1, \ldots, V_{i-1}\right] \leq \exp\left(\frac{\lambda^2}{8}(b_i - a_i)^2\right)$$

(This is because $V_i|V_1, \ldots, V_{i-1}$ is a zero-mean random variable bounded in $[a_i, b_i]$)
Use Chernoff inequality again, $\forall \lambda > 0$

$$\begin{aligned}
\mathbb{P}\left[\frac{1}{n}\sum_{i=1}^{n} V_i > \epsilon\right] &\leq \exp(-\lambda n\epsilon)\mathbb{E}\left[\prod_{i=1}^{n}\exp\left(\lambda V_i\right)\right] \\
&= \exp(-\lambda n\epsilon)\mathbb{E}_{V_1 \ldots V_{n-1}}\left[\mathbb{E}_{V_n}\left[\prod_{i=1}^{n}\exp\left(\lambda V_i\right)|V_1, \ldots, V_{n-1}\right]\right] \\
&= \exp(-\lambda n\epsilon)\mathbb{E}_{V_1 \ldots V_{n-1}}\left[\prod_{i=1}^{n-1}\exp\left(\lambda V_i\right)\mathbb{E}_{V_n}\left[\exp\left(\lambda V_n\right)|V_1, \ldots, V_{n-1}\right]\right] \\
&\leq \exp(-\lambda n\epsilon)\mathbb{E}\left[\prod_{i=1}^{n-1}\exp\left(\lambda V_i\right)\right]\exp\left(\frac{\lambda^2}{8}(b_n - a_n)^2\right) \\
&\leq \ldots \leq \exp(-\lambda^2 n\epsilon)\exp\left(\sum_{i=1}^{n}\frac{\lambda^2}{8}(b_i - a_i)^2\right).
\end{aligned}$$

From this point, the remainder of the proof follows that of the ordinary Hoeffding inequality for sums of independent RVs. In particular, we optimize our bound over $\lambda > 0$, and thereby obtain that

$$\mathbb{P}\left[\frac{1}{n}\sum_{i=1}^{n} V_i > \epsilon\right] \leq \exp\left\{\frac{-2n\epsilon^2}{\frac{1}{n}\sum_{i=1}^{n}(b_i - a_i)^2}\right\},$$

as claimed. □

Reference book: Grimmett & Stirzaker, *Probability & random processes* has further background on martingales and their properties.

13.3 Complexity of function classes

Recap
In proof of classical Glivenko-Cantelli theorem, needed to bound cardinality of

$$\{\mathbb{I}(Z^{(1)} \leq t), \ldots, \mathbb{I}(Z^n \leq t)|t \in \mathbb{R}\}$$

Saw this cardinality is at most $n + 1$.

More generally, let $\mathcal{A}$ be a class of subsets in $\mathbb{R}^d$.
(In previous case, $\mathcal{A} = \mathbb{I}_{left} = \{(-\infty, t]|t \in \mathbb{R}\}$)

Definition

1. For $Z^{(1)}, \ldots, Z^{(n)} \in \mathbb{R}^d$, define

$$S(\mathcal{A}, \{Z^{(i)}\}_{i=1}^n) = card\{A \cap \{Z^{(i)}\}_{i=1}^n | A \in \mathcal{A}\}$$

   Say that $\mathcal{A}$ shatters $\{Z^{(i)}\}_{i=1}^n$ if $S(\mathcal{A}, \{Z^{(i)}\}_{i=1}^n) = 2^n$

2. The nth shatter coefficient is

$$s(\mathcal{A}, n) = \max_{Z^{(1)},\ldots,Z^{(n)}} S(\mathcal{A}, \{Z^{(i)}\}_{i=1}^n)$$

3. The Vapnik-Chernovenkis (VC) dimension of $\mathcal{A}$ is

$$V_{\mathcal{A}} = \sup\{n \in \mathbb{N} | s(\mathcal{A}, n) = 2^n\}$$

Examples:

1. $V_{\mathbb{I}_{left}} = 1$
2. Consider $\mathcal{A} = \{(s, t] | s < t; s, t \in \mathbb{R}\}$. Then

$$\begin{aligned} S(\mathcal{A}, 1) &= 2 \\ S(\mathcal{A}, 2) &= 2^2 \\ S(\mathcal{A}, 3) &= 7 < 2^3 \\ S(\mathcal{A}, n) &= \tfrac{n(n+1)}{2} + 1 \end{aligned}$$

   So $V_{\mathcal{A}} = 2$

Theorem 13.3 (Vapnik-Chernovenkis). *For any class of sets* $\mathcal{A}$

$$\mathbb{P}\left[\sup_{A \in \mathcal{A}} \left|\hat{P}_n(A) - P(A)\right| > \epsilon\right] \leq 8 s(\mathcal{A}, n) \exp\left(-\frac{n\epsilon^2}{32}\right)$$

where $\hat{P}_n(A) = \frac{1}{n}\sum_{i=1}^n \mathbb{I}[Z^{(i)} \in A]$