Spring 2018: STA 6448
Advanced Probability and Inference II
Lecture 4

Yun Yang

- Concentration inequality

Martingales

Definition

A sequence Y_k of random variables adapted to a filtration $\mathcal{F}_k$ is a martingale, if for all k,

$$\mathbb{E}[|Y_k|] < \infty, \quad \text{and} \quad \mathbb{E}[Y_{k+1} \,|\, \mathcal{F}_k] = Y_k.$$

- $\mathcal{F}_k$ is a filtration means these σ-fields are nested: $\mathcal{F}_k \subset \mathcal{F}_{k+1}$.
- Y_k is adapted to $\mathcal{F}_k$ means that each Y_k is measurable w.r.t. $\mathcal{F}_k$.
- If $\mathcal{F}_k = \sigma(X_1, \ldots, X_k)$, the σ-field generated by the first k variables, then Y_k is a martingale sequence w.r.t. X_k.

Martingale difference sequence

Definition

A sequence D_k of random variables adapted to a filtration $\mathcal{F}_k$ is a martingale difference sequence, if for all k,

$$\mathbb{E}[|D_k|] < \infty, \quad \text{and} \quad \mathbb{E}[D_{k+1} \,|\, \mathcal{F}_k] = 0.$$

- For example, $D_k = Y_k - Y_{k-1}$ is a martingale difference sequence
- $Y_k = \sum_{j=0}^{k} D_j$ is a martingale

Example: the Doob construction

Use shorthand $X = (X_1, \ldots, X_n)$ and $X_1^k = (X_1, \ldots, X_k)$.

Define $Y_k = \mathbb{E}[f(X) \mid X_1^k]$ for $k \geq 1$ and $Y_0 = \mathbb{E}[f(X)]$.

Property

If $\mathbb{E}[|f(X)|] < \infty$, then Y_k is a martingale sequence w.r.t. X_k.
Moreover, $D_k = Y_k - Y_{k-1}$ is a martingale difference sequence.

Telescope decomposition:

$$Y_n - Y_0 = \sum_{k=1}^{n} D_k.$$

Example: Likelihood ratio

Let f and g be two density functions, and g is absolutely continuous w.r.t. f.

Suppose X_k are drawn i.i.d. from f, and Y_n is the likelihood ratio,

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

Property

Y_k is a martingale sequence w.r.t. X_k.

Concentration for martingale difference sequences

Theorem

Consider a martingale difference sequence D_k (adapted to a filtration $\mathcal{F}_k$) that satisfies

$$\mathbb{E}\big[\exp(\lambda D_k) \,\big|\, \mathcal{F}_k\big] \leq \exp(\lambda^2 \nu_k^2/2), \quad \textit{a.s. for all } |\lambda| \leq 1/b_k.$$

Then $\sum_{k=1}^n D_k$ is sub-exponential with parameters $(\nu^2, b) = (\sum_{k=1}^n \nu_k^2, \max_k b_k)$, and

$$\mathbb{P}\Big(\Big|\sum_{k=1}^n D_k\Big| \geq t\Big) \leq \begin{cases} 2\exp\left(-\frac{t^2}{2\nu^2}\right) & \textit{if } 0 \leq t \leq \frac{\nu^2}{b}, \\ 2\exp\left(-\frac{t}{2b}\right) & \textit{if } t > \frac{\nu^2}{b}. \end{cases}$$

Proof: Apply the iterative expectation formula.

Concentration for martingale difference sequences

Theorem (Azuma-Hoeffding)

Consider a martingale difference sequence D_k with $|D_k| \leq B_k$ a.s. then

$$\mathbb{P}\Big(\Big|\sum_{k=1}^{n} D_k\Big| \geq t\Big) \leq 2\exp\Big(-\frac{2t^2}{\sum_k B_k^2}\Big).$$

Proof:

$$\mathbb{E}\big[\exp(\lambda D_k) \,\big|\, \mathcal{F}_k\big] \leq \exp(\lambda^2 B_k^2/2) \quad a.s.$$

Bounded difference inequality

Theorem (Bounded difference inequality)

Suppose function $f : \mathbb{R}^n \to \mathbb{R}$ *satisfies the bounded difference property: for all* $x_1, \ldots, x_n, x'_k \in \mathbb{R}$*,*

$$\left| f(x_1, \ldots, x_n) - f(x_1, \ldots, x_{k-1}, x'_k, x_{k+1}, \ldots, x_n) \right| \leq L_k.$$

Then for $X = (X_1, \ldots, X_n)$ *with independent components,*

$$\mathbb{P}\big(|f(X) - \mathbb{E}[f(X)]| \geq t\big) \leq 2 \exp\Big(- \frac{2t^2}{\sum_k L_k^2} \Big).$$

Proof: Apply the Azuma-Hoeffding.

Example: U-statistics

- Let $g : \mathbb{R}^2 \to \mathbb{R}$ be a symmetric function.
- X_k are sequence of i.i.d. random variables.

Definition

Pairwise U-statistics

$$U = \frac{1}{\binom{n}{2}} \sum_{j<k} g(X_j, X_k).$$

For example, if $g(s,t) = |s - t|$, then U is an unbiased estimator of the mean absolute deviation $\mathbb{E}[|X_1 - X_2|]$.

Property

If g is bounded by b, then U is sub-Gaussian with parameter $4b^2/n$.

Example: Rademacher complexity

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

$$Z = \sup_{a \in A} \left(\sum_{k=1}^{n} \varepsilon_k \, a_k \right) = \sup_{a \in A} \langle \varepsilon, \, a \rangle,$$

where $\varepsilon = (\varepsilon_1, \ldots, \varepsilon_n)$ is a sequence of i.i.d. Rademacher variables. Z measures the size of A in a certain sense, and its expectation $\mathcal{R}(A) = \mathbb{E}[Z]$ is known as the Rademacher complexity of set A.

Property

Z is sub-Gaussian with parameter $4 \sum_{k=1}^{n} \sup_{a \in A} a_k^2$.

Apply a deeper result (Talagrand concentration inequality), this sub-Gaussian parameter can be improved to $4 \sup_{a \in A} \sum_{k=1}^{n} a_k^2$.

Lipschitz functions of Gaussian variables

Definition

A function $f: \mathbb{R}^n \to \mathbb{R}$ is L-Lipschitz with respect to the Euclidean norm $\|\cdot\|_2$ if

$$\big|f(x) - f(y)\big| \leq L\,\|x - y\|_2 \qquad \text{for all } x, y \in \mathbb{R}^n.$$

Theorem (Gaussian concentration)

Let $X = (X_1, \ldots, X_n)$ be a vector of i.i.d. standard Gaussian variables, and let $f: \mathbb{R}^n \to \mathbb{R}$ be L-Lipschitz w.r.t. the Euclidean norm. Then $f(X) - \mathbb{E}[f(X)]$ is sub-Gaussian with parameter L, and

$$\mathbb{P}\Big[\big|f(X) - \mathbb{E}[f(X)]\big| \geq t\Big] \leq 2\,e^{-\frac{t^2}{2L^2}} \quad \text{for all } t > 0.$$