36-752 Advanced Statistical Theory **Fall 2018**

Lecture 6: Sept 19

Lecturer: Alessandro Rinaldo *Scribes: Wanshan Li*

Note: *LaTeX template courtesy of UC Berkeley EECS dept.*

6.1 Maximal Inequality

Suppose we have $X_1, \cdots, X_n$ with $\mathbb{E}X_i = 0$ and $X_i \in \text{SG}(\sigma^2)$ for all i. Notice that here X_i's are not necessarily independent! Another thing to keep in mind is that if $X \in \text{SG}(\sigma^2)$ and $\tau^2 > \sigma^2$, then $X_i \in \text{SG}(\tau^2)$.

It is easy to bound

$$\mathbb{P}(\max_i X_i \geq t) \text{ or } \mathbb{P}(\max_i |X_i| \geq t).$$

We can simply use the union bound!

$$\mathbb{P}(\max_i |X_i| \geq t) \leq \sum_{i=1}^n \mathbb{P}(|X_i| \geq t) \leq 2n \exp\left(-\frac{t^2}{2\sigma^2}\right),$$

and to get a high probability bound, we can take $t = \sqrt{2\sigma^2 \log n}$. In the case of Gaussian variables, this maximal inequality is fairly tight, even in constant.

In HW1, we considered

$$\mathbb{P}(\|\hat{\Sigma}_n - \Sigma_n\|_\infty \geq t) \leq \sum_{1\leq i\leq j\leq n} \mathbb{P}(|\hat{\Sigma}_n(i,j) - \Sigma_n(i,j)| \geq t),$$

where there are $d \cdot d/2 = O(d^2)$ terms in the summation.

So far we have upper bounds for $\mathbb{P}(\max_i X_i \geq t)$, and the following theorem provides an upper bound for $\mathbb{E}[\max_i X_i]$.

Theorem 6.1. *Let $X_1, \cdots, X_n$ be random variables such that*

$$\log \mathbb{E}\left[e^{\lambda X_i}\right] \leq \psi(\lambda),\ \forall \lambda \in [0,b),\ 0 < b < \infty,$$

with $\psi(\cdot)$ convex on $[0,b)$. Then

$$\mathbb{E}[\max_i X_i] \leq \inf_{\lambda\in[0,b)} \left\{\frac{\log n + \psi(\lambda)}{\lambda}\right\}.$$

Proof. Suppose $\lambda \in (0,b)$, we have

$$
\begin{aligned}
&\exp\{\lambda\mathbb{E}[\max X_i]\} && \text{[By Jensen's inequality]} \\
\leq &\mathbb{E}\{\exp[\lambda \max X_i]\} = \mathbb{E}\{\max \exp[\lambda X_i]\} && \text{[Monotonicity]} \\
\leq &\sum_{i=1}^{n} \mathbb{E}[\exp(\lambda X_i)] \\
\leq &n \exp(\psi(\lambda)) && \text{[Assumption].}
\end{aligned}
$$

Taking log on both sides and dividing by $\lambda > 0$ complete the proof. □

Example 6.2. *Suppose $X_1, \cdots, X_n \in \mathrm{SG}(\sigma^2)$, then $\log \mathbb{E}\left[e^{\lambda X_i}\right] \leq \psi(\lambda)$ for $\psi(\lambda) = \frac{\lambda^2\sigma^2}{2}$. By Theorem 6.1*

$$
\begin{aligned}
\mathbb{E}\left[\max_{1\leq i\leq n} X_i\right] &\leq \inf_{\lambda>0}\left\{\frac{\log(n) + \frac{\lambda^2\sigma^2}{2}}{\lambda}\right\} \\
&\leq \frac{\log(n) + \frac{2\log(n)}{\sigma^2}\frac{\sigma^2}{2}}{\sqrt{\frac{2\log(n)}{\sigma^2}}} && \left[\textit{Set optimal value } \lambda = \sqrt{\frac{2\log(n)}{\sigma^2}}\right] \\
&= \frac{2\log(n)}{\sqrt{\frac{2\log(n)}{\sigma^2}}} \\
&= \sqrt{2\sigma^2\log(n)}.
\end{aligned}
$$

Briefly, $\mathbb{E}[\max_{1\leq i\leq n} X_i]$ grows on the order of $\sqrt{\log(n)}$.

The following result, Lemma 2.1 in [Ma07], provides an approach to compute $\inf_{\lambda\in[0,b)}\left\{\frac{\log n + \psi(\lambda)}{\lambda}\right\}$.

Proposition 6.3. *If ψ is convex and differentiable on $[0,b)$ with $\psi(0) = \psi'(0) = 0$, which is true if ψ is the logarithm of MGF of a centered RV, then $\forall \mu > 0$,*

$$
\inf_{\lambda\in[0,b)}\left[\frac{\mu + \psi(\lambda)}{\lambda}\right] = \inf\{t \geq 0 : \psi^*(t) \geq \mu\},
$$

where

$$
\psi^*(t) \equiv \sup_{\lambda\in[0,b)}\{\lambda t - \psi(\lambda)\}.
$$

Note The expression $\psi^{*-1}(\mu) := \inf\{t \geq 0 : \psi^*(t) \geq \mu\}$ is called the generalized inversion of ψ^*. For more details, including how to compute $\psi^{*-1}(\mu)$, see [M07] or [BLM13].

Example 6.4. *If $\psi(\lambda) = \frac{\lambda^2\nu^2}{2(1-\lambda b)}$, $\lambda \in [0, 1/b)$, then $\psi^{*-1}(\mu) = \sqrt{2\nu^2\mu} + b\mu$ for $\mu > 0$, thus*

$$
\mathbb{E}[\max_i X_i] \leq \sqrt{2\nu^2 \log n} + b\log n.
$$

Specifically, if $X_i \sim \chi^2_p$, then

$$
\mathbb{E}[\max_i (X_i - p)] \leq 2\sqrt{p\log n} + 2\log n.
$$

6.2 Bounded Difference Inequality

So far we have considered concentration inequalities for $\sum_{i=1}^n X_i$. Suppose now we are interested in $Z = f(X_1, \cdots, X_n)$ here $X_1, \cdots, X_n$ are independent.

Set $Z_0 = \mathbb{E}[f(X_1, \cdots, X_n)]$,

$$Z_k = \mathbb{E}[f(X_1, \cdots, X_n)|X_1, \cdots, X_k], \ k = 1, \cdots, n-1,$$

and $Z_n = f(X_1, \cdots, X_n)$. Then we have

$$f(X_1, \cdots, X_n) - \mathbb{E}[f(X_1, \cdots, X_n)] = Z_n - Z_0 = \sum_{i=1}^n (Z_k - Z_{k-1}) = \sum_{k=1}^n D_k.$$

D_k's are called increments. Before we attack this problem, let's introduce some important tools related to martingales.

Definition 6.5 (Martingale)**.** *Let $\mathcal{F}_0 = \{\emptyset, \Omega\} \subseteq \mathcal{F}_1 \subseteq \cdots \subseteq \mathcal{F}_n \subseteq \cdots$ be a filtration. A sequence of random variables $\{Z_k\}_{k=1,2,\ldots}$ is a martingale if*

1. *Z_k is $\mathcal{F}_k$ measurable;*
2. *$\mathbb{E}[Z_k|\mathcal{F}_{k-1}] = Z_{k-1}$, for $k \geq 2$;*
3. *$\mathbb{E}|Z_k| < \infty$, for all k.*

Example 6.6 (Doob construction)**.** *Consider $Z = f(X_1, \cdots, X_n)$ such that Z is integrable or $\mathbb{E}|Z| < \infty$, and $\mathcal{F}_k = \sigma(X_1, \cdots, X_k)$. Let $Z_k = \mathbb{E}[Z|\mathcal{F}_k]$, then $\{Z_k\}$ is a martingale.*

Example 6.7 (Martingale Difference)**.** *If $(Z_k, \mathcal{F}_k)_{k=0,1,\cdots}$ is a martingale, then the sequence of increments*

$$D_k = Z_k - Z_{k-1}$$

gives a new martingale such that $\mathbb{E}[D_k] = 0$ for all $k \geq 1$. We call $\{D_k\}_{k=1,\ldots}$ a martingale difference.

Theorem 6.8. *Let $\{(D_k, \mathcal{F}_k), k = 1, 2, \cdots,\}$ be a martingale difference s.t.*

$$\mathbb{E}\left[e^{\lambda D_k}|\mathcal{F}_{k-1}\right] \leq e^{\lambda^2 \nu_k^2/2}, \ \forall|\lambda| \leq \frac{1}{\alpha_k}, \tag{6.1}$$

almost surely. Then

1) *$\sum_{k=1}^n D_k \in \mathrm{SE}(\sum_k \nu_k^2, \max_k \alpha_k)$;*

2)

$$\mathbb{P}(|\sum_k D_k| \geq t) \leq \begin{cases} 2\exp\left\{-\frac{t^2}{2\sum_k \nu_k^2}\right\}, \ t \leq \frac{\sum_k \nu_k^2}{\max_k \alpha_k}, \\ 2\exp\left\{-\frac{t}{2\max_k \alpha_k}\right\}, \ t > \frac{\sum_k \nu_k^2}{\max_k \alpha_k}. \end{cases}$$

Proof. 1). By the iterated law of expectation

$$\begin{aligned}\mathbb{E}\left[e^{\lambda\sum_{k=1}^{n}D_k}\right] &= \mathbb{E}\left[\mathbb{E}\left[e^{\lambda\sum_{k=1}^{n}D_k}|\mathcal{F}_{n-1}\right]\right]\\ &= \mathbb{E}\left[\exp\{\lambda\sum_{k=1}^{n-1}D_k\}\mathbb{E}\left[e^{\lambda D_n}|\mathcal{F}_{n-1}\right]\right]\\ &\le \mathbb{E}\left[\exp\{\lambda\sum_{k=1}^{n-1}D_k\}e^{\lambda^2\nu_n^2/2}\right]\\ &= e^{\lambda^2\nu_n^2/2}\mathbb{E}\left[e^{\lambda\sum_{k=1}^{n-1}D_k}\right],\ \text{for}|\lambda|<\frac{1}{\alpha_n},\end{aligned}$$

where we use the fact that $\exp\{\lambda\sum_{k=1}^{n-1}D_k\}\in\mathcal{F}_{n-1}$ and (6.1). Repeating the same procedure for $k=n-1,\cdots,2$, we can get

$$\mathbb{E}\left[e^{\lambda\sum_{k=1}^{n}D_k}\right]\le e^{\lambda^2\frac{\sum_{k=1}^{n}\nu_k^2}{2}},\ \text{for}|\lambda|<\frac{1}{\max_k\alpha_k}.$$

2) Use the property of sub-exponential random variables and 1). □

Corollary 6.9 (Azuma's Inequality or Azuma-Hoeffding Inequality)**.** *Suppose $\{D_k\}_{k=1,2\cdots}$ is a martingale difference. If $D_k\in(a_k,b_k)$ almost surely for some $a_k<b_k$, then*

$$\mathbb{P}\left(\left|\sum_{k=1}^{n}D_k\right|\ge t\right)\le 2\exp\left\{-\frac{2t^2}{\sum_k(b_k-a_k)^2}\right\}.$$

Proof. $D_k\in(a_k,b_k)$ almost surely implies that for almost all $\omega\in\Omega$, the conditional variable $(D_k|\mathcal{F}_{k-1})(\omega)\in(a_k,b_k)$ almost surely, where $(D_k|\mathcal{F}_{k-1})(\omega)$ is defined using regular conditional distributions. By the Hoeffding's bound, $(D_k|\mathcal{F}_{k-1})(\omega)$ is sub-Gaussian with parameter $\sigma^2=(b_k-a_k)^2/4$, for almost all ω. Therefore by the definition of sub-Gaussian r.v. we have that for almost all ω,

$$\mathbb{E}\left[\exp\{\lambda(D_k|\mathcal{F}_{k-1})(\omega)\}\right]\le\exp\left\{\lambda^2\frac{(b_k-a_k)^2}{8}\right\}.$$

By the property of regular conditional distributions (e.g., see [Du2013]),

$$\mathbb{E}\left[e^{\lambda D_k}|\mathcal{F}_{k-1}\right](\omega)=\mathbb{E}\left[\exp\{\lambda(D_k|\mathcal{F}_{k-1})(\omega)\}\right],\text{almost surely.}$$

Therefore

$$\mathbb{E}\left[e^{\lambda D_k}|\mathcal{F}_{k-1}\right]\le\exp\left\{\lambda^2\frac{(b_k-a_k)^2}{8}\right\},\text{almost surely.}$$

Now let $\nu_k^2=(b_k-a_k)^2/4$ and $\alpha_k=0$ in Theorem 6.8 and we can prove the inequality. □

Now we can go back to the original problem, the concentration of $Z=f(X_1,\cdots,X_n)$, where $X_1,\cdots,X_n$ are independent. Briefly speaking, if f is "well behaved", then Z concentrates.

Definition 6.10 (Bounded Difference Property)**.** *A function $f:\mathbb{R}^n\to\mathbb{R}$ satisfies the Bounded Difference Property if $\exists\ L_1,\cdots,L_n$ positive constants such that for all $(x_1,\cdots,x_n)$ in the domain of f and for all $k\in\{1,\cdots,n\}$,*

$$\sup_{x,y}|f(x_1,\cdots,x_{k-1},x,x_{k+1},\cdots,x_n)-f(x_1,\cdots,x_{k-1},y,x_{k+1},\cdots,x_n)|\le L_k.$$

This can be seen as a Lipschitz condition with respect to Hamming distance.

Theorem 6.11 (McDiarmid's Inequality). *Let $X_1, \cdots, X_n$ be independent random variables, $f : \mathbb{R}^n \to \mathbb{R}$ a function that satisfies the Bounded Difference Inequality, with constants $L_1, \cdots, L_n$, and $Z = f(X_1, \cdots, X_n)$. Then*

$$\mathbb{P}(|Z - \mathbb{E}[Z]| \geq t) \leq 2 \exp\left\{ -\frac{2t^2}{\sum_{k=1}^n L_k^2} \right\}.$$

Proof. Recall the Doob construction and let $D_0 = \mathbb{E}[Z] = \mathbb{E}[Z|\mathcal{F}_0]$, $D_k = \mathbb{E}[Z|\mathcal{F}_k]$, for $k = 1, \cdots, n$, where $\mathcal{F}_k = \sigma(X_1, \cdots, X_k)$ and $\mathcal{F}_0 = \{\emptyset, \Omega\}$, then $\{D_k\}_{k=1,2,\cdots}$ is a martingale difference. Moreover, $\sum_{k=1}^n D_k = Z - \mathbb{E}[Z]$. Let

$$A_k = \inf_x \{\mathbb{E}[Z|X_1, \cdots, X_{k-1}, x] - \mathbb{E}[Z|X_1, \cdots, X_{k-1}]\},$$
$$B_k = \sup_x \{\mathbb{E}[Z|X_1, \cdots, X_{k-1}, x] - \mathbb{E}[Z|X_1, \cdots, X_{k-1}]\},$$

for $k = 1, \cdots, n$. Then $D_k \in (A_k, B_k)$ almost surely for all k. By the Bounded Difference Property of f and the independence of $X_1, \cdots, X_n$ we can show that $B_k - A_k \leq L_k$ (see the notes for the next Lecture for details). Apply the Azuma's inequality to $\{D_k\}$ and the result follows. ☐

References

[BLM13] S. Boucheron and G. Lugosi and P. Massart, Concentration Inequalities: a Nonasymptotic Theory of Independence, Oxford University Press, 20

[Du13] R. Durrett, "Probability: Theorey and Examples", Cambridge University Press, 197

[Ma07] D. Massart, Concentration inequalities and model selection, Springer Lecture Notes in Mathematics, vol 1605, 20