36-755: Advanced Statistical Theory **Fall 2017**

Lecture 4: September 13

Lecturer: Alessandro Rinaldo *Scribe: Robin Dunn*

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

Disclaimer: *These notes have not been subjected to the usual scrutiny reserved for formal publications. They may be distributed outside this class only with the permission of the Instructor.*

4.1 Maximal Inequality

Let $X_1, \ldots, X_d$ be centered random variables such that $\log \mathbb{E}\left[e^{\lambda X_i}\right] \leq \psi(\lambda)$ for some convex function $\psi(\cdot)$ and for all λ that satisfy $|\lambda| < \frac{1}{b}$, $b \geq 0$. Then

$$\mathbb{E}\left[\max_{1\leq i\leq d} X_i\right] \leq \inf_{\lambda\in(0,\frac{1}{b})} \left\{\frac{\log(d)+\psi(\lambda)}{\lambda}\right\}.$$

We proved this inequality in the previous class. Here we consider an example.

Suppose $\psi(\lambda) \leq \frac{\lambda^2\sigma^2}{2}$ for all $\lambda \in \mathbb{R}$. This means that $X_i \in SG\left(\sigma^2\right)$. Applying the maximal inequality, we see

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

This tells us that when we have sub-Gaussian $X_1, \ldots, X_d$, $\mathbb{E}\left[\max_{1\leq i\leq d} X_i\right]$ grows on the order of $\sqrt{\log(d)}$.

Also, by the union bound,

$$\mathbb{P}\left(\max_i X_i \geq t\right) \leq \sum_{i=1}^{d} \mathbb{P}(X_i \geq t) \leq de^{-t^2/(2\sigma^2)} = e^{-t^2/(2\sigma^2)+\log(d)}.$$

This probability goes to 0 if $\frac{t^2}{2\sigma^2} \gg \log(d)$.

A maximal inequality for another characterization of the X_is comes from Lemma 2.1 from [M07]:

Lemma 4.1 *Suppose that $X_1, \ldots, X_d$ are centered random variables such that $\log \mathbb{E}\big[e^{\lambda X_i}\big] \le \psi(\lambda)$, $|\lambda| < \frac{1}{b}$, $b \ge 0$ for some function $\psi(\cdot)$ that satisfies*

- *$\psi(\cdot)$ is convex*
- *$\psi(\cdot)$ is continuously differentiable on $[0, \frac{1}{b})$*
- *$\psi(0) = \psi'(0) = 0$.*

Let $\psi^*(t) = \sup_{\lambda \in (0, \frac{1}{b})} \{\lambda t - \psi(\lambda)\}$. Then for all $\mu > 0$, $\psi^{*-1}(\mu) = \inf_{\lambda \in (0, \frac{1}{b})} \left\{ \frac{\mu + \psi(\lambda)}{\lambda} \right\} = \inf\{t \ge 0 : \psi^*(t) > \mu\}$. This expression $\psi^{*-1}(\mu)$ is called the generalized inverse of ψ^*. For more details, see [M07] or [BLM13].

Example: Suppose $X_1, \ldots, X_d$ satisfy the conditions of Lemma 4.1, where $\psi(\lambda) = \frac{\lambda^2 \nu^2}{2(1-\lambda b)}$, $\lambda \in \big(0, \frac{1}{b}\big)$. Then $\psi^{*-1}(\mu) = \sqrt{2\nu^2 \mu} + b\mu$ for $\mu > 0$, and

$$\mathbb{E}\big[\max_i X_i\big] \le \inf_{\lambda \in (0, \frac{1}{b})} \left\{ \frac{\log(d) + \psi(\lambda)}{\lambda} \right\} = \psi^{*-1}(\log(d)) = \sqrt{2\nu^2 \log(d)} + b \log(d).$$

4.2 Bounded Differences

Suppose $X_1, \ldots, X_n$ are independent random variables. So far, most of the concentration inequalities that we have considered have worked with $\sum_{i=1}^n X_i$. More generally, we may be interested in concentration inequalities on arbitrary functions $f(X_1, \ldots, X_n)$. That is, if we let $Z = f(X_1, \ldots, X_n)$, can we place a useful upper bound on $P(|Z - \mathbb{E}(Z)| \ge t)$ for $t > 0$?

We begin by considering the expression $Z - \mathbb{E}(Z)$. Set

$$\begin{aligned}
Z_0 &= \mathbb{E}[f(X_1, \ldots, X_n)] \\
Z_k &= \mathbb{E}[f(X_1, \ldots, X_n) | X_1, \ldots, X_k] \quad \text{for } 1 \le k \le n-1 \\
Z_n &= \mathbb{E}[f(X_1, \ldots, X_n) | X_1, \ldots, X_n] = f(X_1, \ldots, X_n).
\end{aligned}$$

Then we can re-write

$$Z - \mathbb{E}(Z) = Z_n - Z_0 = \sum_{k=1}^n (Z_k - Z_{k-1}) = \sum_{k=1}^n D_k,$$

where $D_k = Z_k - Z_{k-1}$ for $1 \le k \le n$. These terms D_k are not independent, but they are an example of a martingale difference. Martingales can be considered as a first step away from independence. We now turn our attention to martingales.

4.3 Martingales

We begin by defining a martingale.

Definition 4.2 *(Martingale.) Let $(\Omega, \mathcal{F})$ be a measurable space, and let $\mathcal{F}_0 = \{\emptyset, \Omega\}$. Let $\mathcal{F}_0 \subseteq \mathcal{F}_1 \subseteq \mathcal{F}_2 \subseteq \cdots \subseteq \mathcal{F}$ be a sequence of sub-σ-fields. Let $\{Y_k\}_{k=0,1,2,\ldots}$ be a sequence of random variables such that Y_k is $\mathcal{F}_k$-measurable.*

Then the sequence $\{Y_k\}_{k=0,1,2,...}$ *is a martingale adapted to the filtration* $\{\mathcal{F}_k\}_{k=0,1,2,...}$ *if* $\mathbb{E}[|Y_k|] < \infty$ *and* $\mathbb{E}[Y_k|\mathcal{F}_{k-1}] = Y_{k-1}$ *for all* k.

Example: Doob construction.

One way to create a martingale is through the process of Doob construction. Suppose $X_1, \ldots, X_n$ are random variables. Let $Z = f(X_1, \ldots, X_n)$ for some function f, subject to the condition that Z is integrable. Define the generated σ-fields $\mathcal{F}_k = \sigma(X_1, \ldots, X_k)$ for $k \geq 1$. Let $Y_k = \mathbb{E}[Z|\mathcal{F}_k]$. Then the sequence $\{Y_k\}_{k=1,2,...}$ is a martingale.

Proof: We see that $\mathbb{E}[|Y_k|] = \mathbb{E}[|\mathbb{E}[Z|\mathcal{F}_k]|] < \infty$ because Z is integrable. Also, for $k \geq 1$,

$$\mathbb{E}[Y_k|\mathcal{F}_{k-1}] = \mathbb{E}[\mathbb{E}[Z|\mathcal{F}_k]|\mathcal{F}_{k-1}] = \mathbb{E}[Z|\mathcal{F}_{k-1}] = Y_{k-1},$$

where the second equality holds by the Tower Property. We conclude that $\{Y_k\}_{k=1,2,...}$ is a martingale. ■

Exercise: Martingale difference.

Sometimes we work with the difference of consecutive terms of a martingale. Suppose $\{Y_k\}_{k=0,1,2,...}$ is a martingale adapted to the filtration $\{\mathcal{F}_k\}_{k=0,1,2,...}$. Let $\{D_m\}_{m=1,2,...}$ be the sequence defined by $D_m = Y_m - Y_{m-1}$. Then $\mathbb{E}[D_m|\mathcal{F}_{m-1}] = 0$ for all m and $\{D_m\}_{m=1,2,...}$ is adapted to the filtration $\{\mathcal{F}_m\}_{m=1,2,...}$.

Proof: We see that

$$\mathbb{E}[D_m|\mathcal{F}_{m-1}] = \mathbb{E}[Y_m - Y_{m-1}|\mathcal{F}_{m-1}] = \mathbb{E}[Y_m|\mathcal{F}_{m-1}] - \mathbb{E}[Y_{m-1}|\mathcal{F}_{m-1}] = Y_{m-1} - Y_{m-1} = 0.$$

D_m is $\mathcal{F}_m$-measurable because $D_m = Y_m - Y_{m-1}$, and Y_m and Y_{m-1} are both $\mathcal{F}_m$-measurable. ■

Now we relate martingale differences to the concept of sub-exponential variables.

Theorem 4.3 *Let* $\{D_k\}_{k=1,2,...}$ *be a martingale difference with respect to* $\{\mathcal{F}_k\}_{k=1,2,...}$ *such that* $\mathbb{E}\left[e^{\lambda D_k}|\mathcal{F}_{k-1}\right] \leq e^{\lambda^2\nu_k^2/2}$ *a.s. for* $|\lambda| < \frac{1}{\alpha_k}$ *and* $\nu_k, \alpha_k > 0$. *Then*

1. $\sum_{k=1}^n D_k \in SE\left(\sum_{k=1}^n \nu_k^2, \max \alpha_k\right)$

2. *Where* $\nu_*^2 = \displaystyle\sum_{k=1}^n \nu_k^2$, $\alpha_* = \max_k \alpha_k$, *and* $t \geq 0$,

$$\mathbb{P}\left(\left|\sum_{k=1}^n D_k\right| \geq t\right) \leq \begin{cases} 2e^{-t^2/(2\nu_*^2)} & : t \leq \frac{\nu_*^2}{\alpha_*} \\ 2e^{-t/(2\alpha_*)} & : t > \frac{\nu_*^2}{\alpha_*} \end{cases}$$

Proof: To prove statement 1, we see that for $|\lambda| < \frac{1}{\max \alpha_k}$,

$$\begin{aligned}
\mathbb{E}\left[e^{\lambda \sum_{k=1}^n D_k}\right] &= \mathbb{E}\left[\mathbb{E}\left[e^{\lambda \sum_{k=1}^n D_k}|\mathcal{F}_{n-1}\right]\right] && (4.1)\\
&= \mathbb{E}\left[e^{\lambda \sum_{k=1}^{n-1} D_k}\mathbb{E}\left[e^{\lambda D_n}|\mathcal{F}_{n-1}\right]\right] && (4.2)\\
&\leq e^{\lambda^2\nu_n^2/2}\mathbb{E}\left[e^{\lambda \sum_{k=1}^{n-1} D_k}\right] && (4.3)\\
&\leq e^{\lambda^2 \sum_{k=1}^n \nu_k^2/2} && (4.4)
\end{aligned}$$

(4.1) holds by the Tower Property. (4.2) holds because $e^{\lambda \sum_{k=1}^{n-1} D_k}$ is $\mathcal{F}_{n-1}$-measurable. (4.3) holds by the assumptions of Theorem 4.3. (4.4) can be derived by iterating the process of (4.1)-(4.3) $n-1$ more times. This shows that $\sum_{k=1}^n D_k \in SE\left(\sum_{k=1}^n \nu_k^2, \max \alpha_k\right)$.

Statement 2 follows directly by applying the concentration bound on means (sums, in this case) of sub-exponential random variables from the 9-11-17 lecture notes. ■

Corollary 4.4 *(Azuma inequality.) Let $\{D_k\}_{k=1,2,\ldots}$ be a martingale difference with respect to $\{\mathcal{F}_k\}_{k=1,2,\ldots}$. Suppose $a_k \le D_k \le b_k$ a.s. for all k. Then for $t \ge 0$ and $n \in \mathbb{N}$,*

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

Proof: This is a direct consequence of Theorem 4.3. Since $a_k \le D_k \le b_k$ a.s., D_k is sub-Gaussian with parameter at most $\sigma_k^2 = \frac{(b_k-a_k)^2}{4}$. (See properties of sub-Gaussian random variables from 8-30-17 lecture notes.) That implies that $\mathbb{E}\big[e^{\lambda D_k}\big] \le e^{\lambda^2\sigma_k^2/2}$ a.s., so $\mathbb{E}\big[e^{\lambda D_k}|\mathcal{F}_{k-1}\big] \le e^{\lambda^2\sigma_k^2/2}$ a.s. Using statement 2 from Theorem 4.3, we set $\nu_*^2 = \sum_{k=1}^{n}\sigma_k^2 = \frac{1}{4}\sum_{k=1}^{n}(b_k-a_k)^2$ and $\alpha_* = 0$. Then for $t \ge 0$,

$$\mathbb{P}\left(\left|\sum_{k=1}^{n} D_k\right| \ge t\right) \le 2\exp\left\{-\frac{t^2}{2\cdot\frac{1}{4}\sum_{k=1}^{n}(b_k-a_k)^2}\right\} = 2\exp\left\{-\frac{2t^2}{\sum_{k=1}^{n}(b_k-a_k)^2}\right\}.$$

■

4.4 Bounded Differences: The Return of Section 4.2

We return to our problem from Section 4.2, where $X_1, \ldots, X_n$ are independent random variables, $Z = f(X_1, \ldots, X_n)$ for some function f, $Z_k = \mathbb{E}[Z|X_1, \ldots, X_k]$, and D_k is the martingale difference given by $D_k = Z_k - Z_{k-1}$. We want to find a bound on $\mathbb{P}(|Z - \mathbb{E}(Z)| \ge t)$. To make this problem more tractable, we might impose on f the Bounded Difference Property.

Definition 4.5 *(Bounded Difference Property.) A function $f : \mathbb{R}^n \to \mathbb{R}$ satisfies the Bounded Difference Property if for all $x = (x_1, \ldots, x_n) \in \mathbb{R}^n$ in the domain of f and for all $k = 1, \ldots, n$,*

$$\sup_y \Big| f(x_1, \ldots, x_{k-1}, x_k, x_{k+1}, \ldots, x_n) - f(x_1, \ldots, x_{k-1}, y, x_{k+1}, \ldots, x_n)\Big| \le L_k$$

for some positive constants $L_1, \ldots, L_n$. This can be seen as a Lipschitz condition with respect to Hamming distance.

The following theorem uses the Bounded Difference Property to conclude that $Z = f(X_1, \ldots, X_n)$ exhibits sub-Gaussian behavior when f satisfies the Bounded Difference Property.

Theorem 4.6 *(Bounded Difference Inequality, or McDiarmid's Inequality.) Let $(X_1, \ldots, X_n)$ be an n-dimensional random vector with independent components. Suppose $f : \mathbb{R}^n \to \mathbb{R}$ satisfies the Bounded Difference Property with constants $L_1, \ldots, L_n$. Let $Z = f(X_1, \ldots, X_n)$. Then for all $t \ge 0$,*

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

Proof: Recall that we can construct a martingale difference with terms $D_k = \mathbb{E}[Z|X_1, \ldots, X_k] - \mathbb{E}[Z|X_1, \ldots, X_{k-1}]$ for $1 \le k \le n$, and set $D_0 = \mathbb{E}[Z]$. Recall from Section 4.2 that $\sum_{k=1}^{n} D_k = Z - \mathbb{E}[Z]$. For $k = 1, \ldots, n$, define

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

Then $A_k \leq D_k \leq B_k$ a.s. for $k = 1, \dots, n$. We apply the Azuma inequality to show that for $t \geq 0$,

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

since $|B_k - A_k| \leq L_k$ for all k. ■

Example: Density estimation in L_1.

Assumptions:

- $X_1, \dots, X_n \overset{iid}{\sim} P$, where P has Lebesgue density p.
- Let K be a kernel. So $K : \mathbb{R} \to \mathbb{R}_{\geq 0}$ and $\int K(x)dx = 1$.

Our goal is to estimate the density p, which is a function on $\mathbb{R}$.

Define a random function $\hat{p}_h$ by

$$\hat{p}_h(x) = \frac{1}{nh}\sum_{i=1}^n K\left(\frac{X_i - x}{h}\right),$$

where $h > 0$ is the bandwidth. We use $\hat{p}_h$ as an estimator of p. To see why this is a reasonable choice, let $p_h(x) = \mathbb{E}[\hat{p}_h(x)]$ for all $x \in \mathbb{R}$. Then $p_h(x) \geq 0$ and $\int p_h(x)dx = 1$. That means that $p_h(x) = \mathbb{E}[\hat{p}_h(x)]$ is a valid density on $\mathbb{R}$.

The total variation distance between $\hat{p}_h$ and p is defined as $L_1(\hat{p}_h, p) = \int_{\mathbb{R}} |\hat{p}_h(x) - p(x)|\, dx$. We would like to show that $L_1(\hat{p}_h, p) \to 0$, but this is a challenging problem. However, we can at least show that $L_1(\hat{p}_h, p)$ satisfies the Bounded Difference Property.

The total variation distance $L_1(\hat{p}_h, p)$ is a function of the random variables $X_1, \dots, X_n$. Thus, define $L_1(\hat{p}_h, p) = f(x_1, \dots, x_n)$. Define $X^{(1)} = (x_1, \dots, x_{k-1}, x, x_{k+1}, \dots, x_n)$ and $X^{(2)} = (x_1, \dots, x_{k-1}, y, x_{k+1}, \dots, x_n)$. We see that

$$\begin{aligned}
&\Big|f(x_1, \dots, x_{k-1}, x, x_{k+1}, \dots, x_n) - f(x_1, \dots, x_{k-1}, y, x_{k+1}, \dots, x_n)\Big| \\
&\quad = \left|\int_{\mathbb{R}} \left|\frac{1}{nh}\sum_{i=1}^n K\left(\frac{X_i^{(1)} - x}{h}\right) - p(x)\right| dx - \int_{\mathbb{R}} \left|\frac{1}{nh}\sum_{i=1}^n K\left(\frac{X_i^{(2)} - x}{h}\right) - p(x)\right| dx\right| \\
&\quad \leq \frac{1}{nh}\int_{\mathbb{R}} \left|K\left(\frac{x - z}{h}\right) - K\left(\frac{y - z}{h}\right)\right| dz \\
&\quad \leq \frac{1}{nh}\left[\int_{\mathbb{R}} K\left(\frac{x - z}{h}\right) dz + \int_{\mathbb{R}} K\left(\frac{y - z}{h}\right) dz\right]
\end{aligned}$$

Setting $w = \frac{x-z}{h}$ and $w' = \frac{y-z}{h}$,

$$= \frac{1}{nh}\left[h\int_{\mathbb{R}} K(w)dw + h\int_{\mathbb{R}} K(w')dw'\right]$$

Since K is a density, both integrals equal 1, so this final expression equals $\frac{2}{n}$.

This shows that $L_1(\hat{p}_h, p)$ satisfies the bounded difference property with constant $\frac{2}{n}$ for each of the n components. Applying the Bounded Difference Inequality (McDiarmid's Inequality), we determine that for $t \geq 0$,

$$\mathbb{P}\left(\left|L_1(\hat{p}_h, p) - \mathbb{E}[L_1(\hat{p}_h, p)]\right| \geq t\right) \leq 2\exp\left\{-\frac{2t^2}{n\left(\frac{2}{n}\right)^2}\right\} = 2\exp\left\{-\frac{nt^2}{2}\right\}.$$

This bound does not depend on the bandwidth h.

Example: Uniform deviation.

Let $X_1, \ldots, X_n \overset{iid}{\sim} P$ in $\mathbb{R}^d$. Let $\mathcal{A}$ be a collection of subsets in $\mathbb{R}^d$. Construct an empirical measure $P_n(B) = \dfrac{1}{n}\displaystyle\sum_{i=1}^n I\{X_i \in B\}$, where the sets B are Borel. Often we are interested in $\sup\limits_{A\in\mathcal{A}} |P(A) - P_n(A)|$. As an example, let $d = 1$ and $\mathcal{A} = \{(-\infty, x] : x \in \mathbb{R}\}$. Then $\sup\limits_{A\in\mathcal{A}} |P(A) - P_n(A)| = \sup\limits_{x} |F(x) - F_n(x)|$, where $F(x)$ is the CDF of P and $F_n(x)$ is the empirical CDF.

$P_n(A)$ satisfies the Bounded Difference Property with constants $\frac{1}{n}$ because changing one of the X_is will change $P_n(A)$ by at most $\frac{1}{n}$. So changing one of the X_is will change $\sup\limits_{A\in\mathcal{A}} |P(A) - P_n(A)|$ by at most $\frac{1}{n}$. Applying the Bounded Difference Inequality,

$$\mathbb{P}\left(\left|\sup_{A\in\mathcal{A}} |P(A) - P_n(A)| - \mathbb{E}\big[\sup_{A\in\mathcal{A}} |P(A) - P_n(A)|\big]\right| \geq t\right) \leq 2\exp\left\{-\frac{2t^2}{\sum_{i=1}^n \left(\frac{1}{n}\right)^2}\right\} = 2e^{-2t^2 n}.$$

The choice of set $\mathcal{A}$ is nowhere in this bound.

References

[BLM13] S. Boucheron and G. Lugosi and P. Massart, "Concentration inequalities: a nonasymptotic theory of independence," *Oxford University Press*, 2013.

[M07] D. Massart, "Concentration inequalities and model selection," *Springer Lecture Notes in Mathematics*, vol 1605, 2007.