36-755: Advanced Statistical Theory I **Fall 2016**

Lecture 5: September 14

Lecturer: Alessandro Rinaldo *Scribes: Ilmun Kim*

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

5.1 Martingale-based methods

Last time, we studied tail bounds on the maximum of random variables as well as a quadratic form of random variables. Now we turn our attention to concentration inequalities of more general functions.

5.1.1 Bounded difference inequality

Theorem 5.1 *Let $\{D_k, \mathcal{F}_k\}_{k=1}^\infty$ be a martingale difference sequence and suppose that $\mathbb{E}\left[e^{\lambda D_k} | \mathcal{F}_{k-1}\right] \leq e^{\lambda^2 \nu_k^2/2}$ almost everywhere (a.e.) for any $|\lambda| < 1/\alpha_k$ and ν_k, $\alpha_k > 0$. Then $\sum_{k=1}^n D_k$ is sub-exponential with parameters $(\sqrt{\sum_{k=1}^n \nu_k^2}, \alpha_*)$.*

Proof: For $\lambda \in (-1/\alpha_*, 1/\alpha_*)$, apply iterated expectation to get

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

which proves the result. ■

The sub-exponential tail bound provides the following inequality.

Corollary 5.2

$$\mathbb{P}\left[|\sum_{k=1}^n D_k| \geq t\right] \leq \begin{cases} 2e^{-\frac{t^2}{2\sum_{k=1}^n \nu_k^2}} & \text{if } 0 \leq t \leq \frac{\sum_{k=1}^n \nu_k^2}{\alpha_*} \\ 2e^{-\frac{t}{2\alpha_*}} & \text{if } t > \frac{\sum_{k=1}^n \nu_k^2}{\alpha_*} \end{cases}$$

Remember that bounded random variables are sub-Gaussian, which gives the following corollary.

Corollary 5.3 [Azuma-Hoeffding] *Let $\{D_k, \mathcal{F}_k\}_{k=1}^\infty$ be a martingale difference sequence such that $D_k \in [a_k, b_k]$ almost surely for all $k = 1, \ldots, n$. Then for all $t > 0$,*

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

Proof: Since $D_k \in [a_k, b_k]$ almost surely, the conditioned variable $(D_k|F_{k-1})$ is also bounded in $[a_k, b_k]$ almost surely. Therefore, $(D_k|F_{k-1})$ is sub-Gaussian at most $\sigma = (b_k - a_k)/2$ for all $k = 1, \dots, n$. The result follows by Theorem 5.1 and Corollary 5.2 with parameters $(\sqrt{\sum_{k=1}^n (b_k - a_k)^2/4}, 0)$. ■

As an application of these results, we will establish a useful inequality, which is called the *bounded difference inequality* or *McDiarmid's inequality*. Let us begin by defining the bounded difference property.

Definition 5.4 [Bounded difference property] *A function $f : \mathbb{R}^d \to \mathbb{R}$ satisfies the bounded difference property (BDP) if there exists positive constants $(L_1, \dots, L_n)$ such that for each $k = 1, 2, \dots, n$,*

$$|f(x_1, \dots, x_{k-1}, x, x_{k+1}, \dots, x_n) - f(x_1, \dots, x_{k-1}, x', x_{k+1}, \dots, x_n)| \le L_i \quad \textit{for all} \;\; x, x' \in \mathbb{R}^d.$$

Theorem 5.5 [Bounded difference inequality] *Suppose that $Z = f(X)$ satisfies the bounded difference property with parameters $(L_1, \dots, L_n)$ and that the random vector $X = (X_1, \dots, X_n)$ has independent elements. Then*

$$\mathbb{P}(|Z - \mathbb{E}(Z)| \ge t) \le 2 \exp\left(-\frac{2t^2}{\sum_{k=1}^n L_k^2}\right) \quad \textit{for all } t \ge 0.$$

Proof: Start by constructing a martingale difference using the Doob martingale decomposition of Z as

$$\begin{aligned} D_0 &= \mathbb{E}(Z) \\ D_k &= \mathbb{E}(Z|X_1, \dots, X_k) - \mathbb{E}(Z|X_1, \dots, X_{k-1}) \quad \text{for } k = 1, \dots, n. \end{aligned}$$

Then we have $Z - \mathbb{E}(Z) = \sum_{k=1}^n D_k$. Define the random variables

$$\begin{aligned} A_k &= \inf_x \mathbb{E}(Z|X_1, \dots X_{k-1}, x) - \mathbb{E}(Z|X_1, \dots, X_{k-1}) \quad \text{and} \\ B_k &= \sup_x \mathbb{E}(Z|X_1, \dots, X_{k-1}, x) - \mathbb{E}(Z|X_1, \dots, X_{k-1}) \end{aligned}$$

so that $B_k \ge A_k$ a.e. for all $k = 1, \dots, n$. In addition,

$$D_k - A_k = \mathbb{E}(Z|X_1, \dots, X_k) - \inf_x E(Z|X_1, \dots, X_{k-1}, x) \ge 0 \quad \text{a.e.}$$

Similarly, $B_k - D_k \ge 0$ a.e. Now observe that

$$\begin{aligned} D_k &\le B_k - A_k \\ &\le \sup_{x,x'} \Big| \mathbb{E}\left[Z|X_1, \dots, X_{k-1}, x\right] - \mathbb{E}\left[Z|X_1, \dots, X_{k-1}, y\right] \Big| \\ &\le L_k. \end{aligned}$$

Apply the Azuma-Hoeffding inequality to get the result. ■

5.1.2 Applications

Example 5.6 [Kernel density estimate] *Let $X_1, \dots, X_n$ be independent and identically distributed random samples from a distribution P with a Lebesgue-density $p = dP/d\mu$. We are interested in estimating the shape of p. Its kernel density estimate is*

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

where $K(x) \geq 0$, $\int K(x)dx = 1$ *and* $h > 0$. *One way of measuring a proximity between* $\hat{p}_h$ *and* p *is*

$$Z = \int_{-\infty}^{\infty} |\hat{p}_h(x) - p(x)|dx = f(X_1, \ldots, X_n).$$

Then, denote $\hat{p}_h'(x)$ *for the kernel density estimate obtained by replacing* X_i *by* X_i' *and bound*

$$\begin{aligned}\Big|f(X_1, \ldots, X_i', \ldots, X_n) - f(X_1, \ldots, X_i, \ldots, X_n)\Big| &= \Big|\int_{-\infty}^{\infty} |\hat{p}_h'(x) - p(x)|dx - \int_{-\infty}^{\infty} |\hat{p}_h(x) - p(x)|dx\Big| \\ &\leq \frac{1}{nh}\int_{-\infty}^{\infty}\Big|K\left(\frac{x - X_i'}{h}\right) - K\left(\frac{x - X_i}{h}\right)\Big|dx \\ &\leq \frac{1}{nh}\left[h\int_{-\infty}^{\infty} K(z')dz' + h\int_{-\infty}^{\infty} K(z)dz\right] = \frac{2}{n}\end{aligned}$$

where we used the triangle inequality and the variable transformation to get the bound. This shows that f *satisfies the bounded difference property with* $L_k = 2/n$ *for all* $k = 1, \ldots, n$. *Then McDiarmids inequality gives*

$$\mathbb{P}(|Z - \mathbb{E}(Z)| \geq t) \leq 2\exp(-\frac{nt^2}{2})$$

where the upper bound does not depend on h.

Example 5.7 [Empirical measure] *Let* $\mathcal{A}$ *be a class of sets in* $\mathbb{R}^d$ *and* $X_1, \ldots, X_n$ *be independent and identically distributed random samples from a distribution* $\mathbb{P}$ *on* $\mathbb{R}^d$. *We are interested in*

$$Z = \sup_{A \in \mathcal{A}} \big|\mathbb{P}(A) - \mathbb{P}_n(A)\big|$$

where $\mathbb{P}_n(A) = \frac{1}{n}\sum_{i=1}^n I(X_i \in A)$ *is the empirical measure of* A. *The empirical distribution function provides an example of empirical measures when* $d = 1$. *For a class* $\mathcal{A} = \{(-\infty, x] : x \in \mathbb{R}\}$,

$$Z_1 = \sup_t \big|F_n(t) - F(t)\big|$$

where $F_n(t) = \mathbb{P}_n((-\infty, t])$ *and* $F(t) = \mathbb{P}(X \leq t)$. *In particular, Glivenko-Cantelli theorem says that* $Z_1 \to 0$ *almost surely. Later on, we will look into bounds on* Z. *For now, denote* $Z = f(X_1, \ldots, X_n)$ *and* $\mathbb{P}_n'(A)$ *for the empirical measure of* A *obtained by replacing* X_i *by* X_i'

$$\begin{aligned}\big|f(X_1, \ldots, X_i', \ldots, X_n) - f(X_1, \ldots, X_i, \ldots, X_n)\big| &= \big|\sup_{A \in \mathcal{A}}\big|\mathbb{P}(A) - \mathbb{P}_n'(A)\big| - \sup_{A \in \mathcal{A}}\big|\mathbb{P}(A) - \mathbb{P}_n(A)\big|\big| \\ &\leq \sup_{A \in \mathcal{A}}\big|\mathbb{P}_n'(A) - \mathbb{P}_n(A)\big| = \frac{1}{n}.\end{aligned}$$

Hence, Z *satisfies the bounded difference property with* $L_k = 1/n$ *for all* $k = 1, \ldots, n$. *Then McDiarmid's inequality provides*

$$\mathbb{P}(|Z - \mathbb{E}(Z)| \geq t) \leq 2\exp(-2nt^2).$$

5.2 Lipschitz functions of Gaussian variables

We investigate the concentration properties of Lipschitz functions of Gaussian variables. Let us say that a function $f : \mathbb{R}^d \to \mathbb{R}$ is L-Lipschitz with respect to the Euclidean norm $||\cdot||_2$ if

$$|f(x) - f(y)| \leq L||x - y||_2 \quad \text{for } x, y \in \mathbb{R}^d.$$

A Lipschitz function is absolutely continuous and thus is differentiable almost everywhere. Now, the following theorem guarantees that any Lipschitz function of Gaussian variables is sub-Gaussian with parameter at most L.

Theorem 5.8 *Let $(X_1, \ldots, X_n)$ be a vector of i.i.d. Gaussian variables from $N(0, \sigma^2)$ and let $f : \mathbb{R}^d \to \mathbb{R}$ be L-Lipschitz. Then the variable $f(X) - \mathbb{E}(f(X))$ is sub-Gaussian with parameter at most L, and thus*

$$\mathbb{P}\left[\left|f(X) - \mathbb{E}(f(X))\right| \geq t\right] \leq 2\exp\left(-\frac{t^2}{2L^2\sigma^2}\right) \quad \text{for all } t \geq 0.$$

Remarkably, this is a dimension free inequality.

Proof: Refer to [BLM13] in p.125. ■

Example 5.9 [Maximum of Gaussian variables] *For a random vector $X = (X_1, \ldots, X_d) \sim N_d(0, \Sigma)$, define $Z = \max_{1\leq i\leq d} X_i$ or $Z = \max_{1\leq i\leq d} |X_i|$ and $\sigma^2_{max} = \max_{1\leq i,j\leq d} \Sigma_{i,j}$. Then,*

$$\mathbb{P}\left[|Z - \mathbb{E}(Z)| \geq t\right] \leq 2\exp\left(-\frac{t^2}{2\sigma^2_{max}}\right).$$

Proof: *Denote $X = AW$ where $W \sim N_d(0, I)$ and $AA^T = \Sigma$. Then $Z = \max_{1\leq i\leq d} X_i = f(W)$ where $f : \mathbb{R}^d \to \mathbb{R}$ is the function*

$$f(x) = \max_{1\leq i\leq d} (Ax)_i$$

Notice that the function f is Lipschitz with the parameter $L = \max_{1\leq i\leq d} \sqrt{\sum_{j=1}^d A^2_{i,j}}$ because for $x, y \in \mathbb{R}^d$ we have

$$\begin{aligned} \left|(Ax)_i - (Ay)_i\right| &= \Big|\sum_{j=1}^d A_{i,j}(x_j - y_j)\Big| \\ &\leq \sqrt{\sum_{j=1}^d A^2_{i,j}}||x - y||_2 \end{aligned}$$

by Cauchy–Schwarz inequality. Furthermore,

$$\sum_{j=1}^d A^2_{i,j} = \mathbb{V}(X_i) = \mathbb{V}\left[\sum_{j=1}^d A_{i,j} Z_j\right].$$

Therefore, f is σ_{max}-Lipschitz. The proof is done by Theorem 5.8. ■

5.3 Covering and packing number

Let Y_i be X_i or $|X_i|$ where X_i is sub-Gaussian or sub-Exponential. In this case, we are often interested in $\max_{i\in\mathcal{I}} Y_i$ or $\mathbb{E}\left[\max_{i\in\mathcal{I}} Y_i\right]$ for a given class $\mathcal{I}$. If the size of $\mathcal{I}$ is infinite, it is challenging to develop uniform bounds. To tackle this problem, we will discretize $\mathcal{I}$ by picking a finite subset $\tilde{\mathcal{I}}$ of $\mathcal{I}$ and then approximating $\max_{i\in\mathcal{I}} Y_i$ with $\max_{i\in\tilde{\mathcal{I}}} Y_i$. Before we go into the details, let us define a metric space.

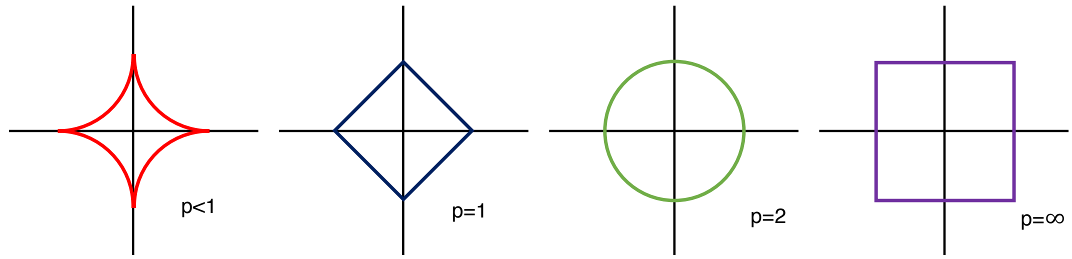

Figure 5.1: Unit sphere in L_p

Definition 5.10 [Metric space] *A metric space is an ordered pair* $(\mathcal{X}, d)$ *where* $\mathcal{X}$ *is a set and* d *is a metric on* $\mathcal{X}$ *such that* $d : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ *and for* $x, y, z \in \mathcal{X}$*, the following holds.*

1. $d(x, y) \geq 0$ *and* $d(x, y) = 0$ *if and only if* $x = y$
2. $d(x, y) = d(y, x)$
3. $d(x, z) \leq d(x, y) + d(y, z)$

Example 5.11 *Here are some examples of metric spaces.*

- $(\mathbb{R}^d, ||\cdot||)$ *and* $||x|| = \sqrt{\sum_i x_i^2}$
- $(\mathbb{R}^d, ||\cdot||_p)$ *and* $||x||_p = (\sum_i x_i^p)^{1/p}$ *for* $p \geq 1$
- $(\mathbb{R}^d, ||\cdot||_\infty)$ *and* $||x||_\infty = \max_i |x_i|$
- $(\{0,1\}^d, d_H)$ *and* $d_H(x, y) = \frac{1}{d}\sum_{i=1}^d I(x_i \neq y_i)$ *called the Hamming distance*

Lastly, we talked about L_p-space. Let $\mathcal{X} = \{f : [0,1] \to \mathbb{R}\}$ be a set of functions. An L_p-space on $[0,1]$ contains functions of $\mathcal{X}$ for which the p-th power of the absolute value is μ-integrable. That is

$$||f||_p = \left(\int_0^1 |f|^p d\mu\right)^{1/p} < \infty$$

where μ is a measure on $[0,1]$ and $p \geq 1$. The common choice of p is $p = 2$, which allows a richer theory. The L_p-distance between f and g is defined as

$$||f - g||_p = \left(\int_0^1 |f(x) - g(x)|^p d\mu\right)^{\frac{1}{p}}.$$

Especially, when $p = \infty$,

$$||f - g||_\infty = \sup_{x\in[0,1]} |f(x) - g(x)|.$$

References

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