Lecture 7: September 24

Lecturer: Alessandro Rinaldo *Scribes: Maria Jahja*

7.1 Bounded Differences Inequality

Continuing from the previous lecture, we now show the proof of the *bounded differences inequality*, also known as McDiarmid's inequality.

Theorem 7.1 (Bounded Differences Inequality) *Suppose $(X_1, \ldots, X_n)$ are independent random variables, and let $f : \mathbb{R}^n \to \mathbb{R}$ satisfy the bounded differences property with constants $L_1, \ldots, L_n$.*

Then

$$\mathbb{P}(|Z - \mathbb{E}[Z]| \geq t) \leq 2 \exp\left(-\frac{2t^2}{\sum_i^n L_i^2}\right)$$

where $Z = f(X_1, \ldots, X_n)$.

Proof: Define the martingale difference

$$D_k = \mathbb{E}[Z|X_1, \ldots, X_k] - \mathbb{E}[Z|X_1, \ldots, X_{k-1}]$$

for $k = 1, \ldots, n$ and $D_0 = \mathbb{E}[Z]$. Then we have $Z - \mathbb{E}[Z] = \sum_i^n D_k$. If we also define

$$\begin{aligned}
A_k &= \inf_x \mathbb{E}[Z|X_1, \ldots, X_{k-1}, x] - \mathbb{E}[Z|X_1, \ldots, X_{k-1}] \\
&= \inf_x \int f(X_1, \ldots, X_{k-1}, x, x_{k+1}, \ldots, x_n) dP(x_{k+1}) \cdots dP(x_n) \\
B_k &= \sup_x \mathbb{E}[Z|X_1, \ldots, X_{k-1}, x] - \mathbb{E}[Z|X_1, \ldots, X_{k-1}] \\
&= \sup_x \int f(X_1, \ldots, X_{k-1}, x, x_{k+1}, \ldots, x_n) dP(x_{k+1}) \cdots dP(x_n),
\end{aligned}$$

then we have sandwiched A_k

$$A_k \leq D_k \leq B_k \text{ a.e } \forall k = 1, \ldots, n.$$

We need to bound the quantity $B_k - A_k$. By independence of the X_k and the bounded difference assumption

$$\begin{aligned}
B_k - A_k &= \sup_x \mathbb{E}[Z|X_1, \ldots, X_{k-1}, x] - \inf_x \mathbb{E}[Z|X_1, \ldots, X_{k-1}, x] \\
&= \sup_{x,y} \int f(X_1, \ldots, X_{k-1}, x, x_{k+1}, \ldots, x_n) \\
&\qquad - f(X_1, \ldots, X_{k-1}, y, x_{k+1}, \ldots, x_n) dP(x_{k+1}) \cdots dP(x_n) \\
&\leq L_k.
\end{aligned}$$

We apply Azuma-Hoeffding (as proven in the last lecture), and the result follows. ■

7.1.0.1 Examples of Bounded Differences Inequality

1. **U-statistics**. Let $X_1, \ldots, X_n \overset{\text{iid}}{\sim} P$, and let $g : \mathbb{R}^2 \to \mathbb{R}$ such that g is symmetric in its arguments. A U-statistic of order 2 is a random variable with the form

$$U_n = \frac{1}{\binom{n}{2}} \sum_{i<j} g(X_i, X_j).$$

- One example is $g(x_1, x_2) = \frac{1}{2}(x_1 - x_2)^2$. Then $\mathbb{E}[g(X_1, X_2)] = \text{Var}[X_1]$.
- Another example: $\mu = \mathbb{P}(X_1 + X_2 \geq 0)$. If P is symmetric around 0, then $\mu = 1/2$.
- A third example (Kendall's tao): let $Z_i = (X_i, Y_i) \overset{iid}{\sim} P$, for $i = 1, \ldots, n$. Let

$$\tau = \frac{4}{n(n-1)} \sum_{i<j} \mathbf{1}\{(Y_j - Y_i)(X_j - X_i) > 0\} - 1$$

which calculates the fraction of concordant pairs, where we have concordance if $(Y_j - Y_i)(X_j - X_i) > 0$. Then $\tau + 1$ is a U-statistic with order 2 given by

$$g\left(\begin{pmatrix} x_1 \\ x_2 \end{pmatrix}, \begin{pmatrix} y_1 \\ y_2 \end{pmatrix}\right) = 2 \times \mathbf{1}\{(y_2 - y_1)(x_2 - x_1) > 0\}$$

If $X \perp Y$ (and both have continuous distributions), then $\mathbb{E}[\tau] = 0$.
- A fact is that if U_n is a U-statistic such that $\mathbb{E}[U_n] = \theta$, where θ is some parameter of interest, then

$$\text{Var}[U_n] \leq \text{Var}[T],$$

where T is any unbiased estimator of θ.

What is the concentration of U_n around $\mathbb{E}[X_n]$? Let's further assume that the U-statistic kernel g is bounded in L^∞-norm, i.e.

$$||g||_\infty = \sup_{x \in \mathbb{R}^2} |g(x)| \leq b.$$

We can check the bounded difference property, first expressing U_n as $U_n = f(x_1, \ldots, x_n)$. Then for all $x_1, \ldots, x_n$, and $(x, y) \in \mathbb{R}$

$$\begin{aligned} |f(x_1, \ldots, x_{k-1}, x, x_{k+1}, \ldots, x_n) - f(x_1, \ldots, x_{k-1}, y, x_{k+1}, \ldots, x_n)| \\ \leq \frac{1}{\binom{n}{2}} \sum_{j \neq k} |g(x, x_j), g(y, x_j)| \\ \leq \frac{(n-1)}{\binom{n}{2}} 2b = \frac{4b}{n}. \end{aligned}$$

By the bounded differences inequality,

$$\mathbb{P}(|U_n - \mathbb{E}[U_n]| \geq t) \leq 2 \exp\left(-\frac{nt^2}{8b^2}\right).$$

In general, for a U-statistic of order m with the form

$$U_n = \frac{1}{\binom{n}{m}} \sum_{i_1 < i_2 < \cdots < i_m} g(X_{i_1}, \ldots, X_{i_m}),$$

we can form a bound

$$\mathbb{P}(|U_n - \mathbb{E}[U_n]| \geq t) \leq 2\exp\left(-\frac{nt^2}{2m^2b^2}\right).$$

Note that better bounds exist of the order $\exp\left(-\frac{nt^2}{m}\right)$ (Hoeffding, 1948), although this is a simple way to start.

2. **Clique number in Erdös-Renyi random graphs**. Let $G = \{X_{i,j}\}_{i<j}$ be a random graph with n vertices, where $X_{i,j} \overset{\text{iid}}{\sim} \text{Bernoulli}(p),\ p \in (0,1)$. Define C as the *clique number*, or the size of the largest complete subgraph. What can we say about the clique number? The bounded difference inequality gives a bound of the form

$$\mathbb{P}\left(\left|\frac{C}{n} - \frac{\mathbb{E}[C]}{n}\right| \geq t\right) \leq 2\exp(-2nt^2).$$

However, there is one problem: What is $\mathbb{E}[C]$?

3. **Empirical measure**. Let $\mathcal{A}$ be a collection of subsets in $\mathbb{R}^d$, and let $X_1, \ldots, X_n \overset{\text{iid}}{\sim} P$ on $(\mathbb{R}^d, \mathcal{B}_n)$. For each $A \in \mathcal{A}$, define the empirical measure as

$$P_n(A) = \frac{1}{n}\sum_{i=1}^n \mathbf{1}\{X_i \in A\},$$

therefore $\mathbb{E}[P_n(A)] = P(A)$ for all A. We are interested in the largest deviation to the true measure

$$Z \triangleq \sup_{A \in \mathcal{A}} |P_n(A) - P(A)|.$$

For example, take $d = 1, \mathcal{A} = \{(-\infty, x],\ x \in \mathbb{R}\}$. Then

$$Z = \sup_{A \in \mathcal{A}} |P_n(A) - P(A)| = \sup_x |F_n(x) - F(x)|,$$

which is the empirical CDF. Then by the bounded difference inequality,

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

But, note we still have the same issue: What is $\mathbb{E}[Z]$? This is the next topic once we cover a few more useful concentration inequalities.

7.2 Concentration Inequalities

We will cover several noteworthy concentration equalities. The results are stated below without proof–see Chapters 2 and 3 of Wainwright for details.

7.2.1 Lipschitz Functions of Gaussians

Let $Z_1, \ldots Z_n \overset{\text{iid}}{\sim} \mathcal{N}(0,1)$, and let $f : \mathbb{R}^n \to \mathbb{R}$ be L-Lipschitz, i.e.

$$|f(x) - f(y)| \leq L||x - y||.$$

Then,

$$\mathbb{P}(|f(z_1, \ldots, z_n) - \mathbb{E}[f(z_1, \ldots, z_n)]| \geq t) \leq 2\exp\left(-\frac{t^2}{2L^2\sigma^2}\right).$$

What is striking about this result is that the dimension does not appear in the bound! This is a *dimension-free* bound, as long as we keep the Lipschitz assumption.

Corollary 7.2 *Let* $Y \sim \mathcal{N}_d(0, \Sigma)$ *and let*

$$X = \max_i Y_i \quad or \quad \max_i |Y_i|.$$

Then

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

where $\sigma^2 = \max_i \Sigma_{ii}$.

7.2.2 Maximal Inequalities

Often, we are interested in computing high probability bounds and the expected values of the quantities

$$\sup_{i \in \mathcal{I}} X_i \quad \text{or} \quad \sup_{i \in \mathcal{I}} |X_i|$$

where $\mathcal{I}$ is some (possibly infinite) set. If $\mathcal{I}$ is finite, we can find bounds through union bounds or some other properties (sub-Gaussianity) of the random variables. But what about infinite and uncountable $\mathcal{I}$?

We can first try to approximate the set with a finite subset. As a first approach, consider a discretization of the set by evaluating only a grid of points over the space.

7.2.2.1 Approximating large spaces

First, recall the definition of a metric space.

Definition 7.3 (Metric space) *A metric space is a pair* $(\mathcal{X}, d)$, *where* $\mathcal{X}$ *is an arbitrary set, and* d *is a metric,* $d : \mathcal{X} \times \mathcal{X} \to [0, \infty)]$ *with the following properties for any* $x, y, z \in \mathcal{X}$

1. $d(x, y) \geq 0$ *(Non-negativity)*
2. $d(x, y) = 0 \Leftrightarrow x = y$ *(Identity of indiscernibles)*
3. $d(x, y) = d(y, x)$ *(Symmetry)*
4. $d(x, z) \leq d(x, y) + d(y, z)$ *(Triangle inequality)*

Examples

- The set $\mathbb{R}^d$ with the L^p-norm defined for $1 \leq p \leq \infty$ is a d-dimensional normed vector space. This is the pair $(\mathbb{R}^d, ||\cdot||_p)$, where $||x||_p = \left(\sum_i |x_i|^p\right)^{1/p}$. If $p = \infty$, take $||x||_\infty = \max_i |x_i|$.
- Consider a discrete space where $\mathcal{X} = \{0, 1\}^d$, and d is the normalized Hamming's distance

$$d_H(x, y) = \frac{1}{n} \sum_{i=1}^{d} \mathbf{1}(x_i \neq y_i).$$

- (L^p spaces). Let $\mathcal{X}$ be the set of real-valued functions on $[0, 1]$, and

$$d_p(f, g) = ||f - g||_p = \left(\int_0^1 |f(x) - g(x)|^p \; dx \right)^{1/p}.$$

  Note that $L^p(\mathcal{X}, d)$ consists of equivalence classes. If we have $p = \infty$, then

$$||f - g||_\infty = \sup_x |f(x) - g(x)|,$$

  which is a metric on $C([0, 1])$, or the set of continuous functions on $[0, 1]$.

Next class, we will study covering and packing numbers.