Math 547: Mathematical Foundations of Statistical Learning Theory

Lectures — Week 6

Instructor: Stas Minsker *Scribe: S. Minsker*

Let $X_1, ...X_n$ be a sequence of random variables. How to estimate $\mathbb{E}\left[\max_{j=1,\dots,n} X_j\right]$?

Proposition 1. Let $X_1 \in \mathrm{SG}(\sigma_1^2)$,...,$X_n \in \mathrm{SG}(\sigma_n^2)$ (not necessarily independent!). Then

$$\mathbb{E}\max_{j=1,\dots,n} X_j \le \sqrt{2}\sigma\sqrt{\log n},$$

where $\sigma = \max\left(\sigma_1, ..., \sigma_n\right)$.

Proof. Let $\lambda > 0$. Applying Jensen's inequality to the function $f(x) = e^{\lambda x}$, we get

$$\begin{aligned}\exp\left(\lambda\mathbb{E}\max_{j=1,\dots,n} X_j\right) &\le \mathbb{E}\exp\left(\lambda\max_{j=1,\dots,n} X_j\right) = \mathbb{E}\max_{j=1,\dots,n} e^{\lambda X_j} \\ &\le \mathbb{E}\sum_{j=1}^n e^{\lambda X_j} \le \sum_{j=1}^n e^{\frac{\lambda^2\sigma_j^2}{2}} \le n e^{\frac{\lambda^2\sigma^2}{2}}\end{aligned}$$

Taking logarithm on both sides of the inequality, we obtain

$$\mathbb{E}\max_{j=1,\dots,n} X_j \le \frac{\log n}{\lambda} + \lambda\frac{\sigma^2}{2}.$$

Minimizing the right-hand side over λ results in the bound $\mathbb{E}\max_{j=1,\dots,n} X_j \le \sqrt{2\log n}\cdot\sigma$. □

Corollary 1. $\mathbb{E}\max_{j=1,\dots,n}|X_j| \le \sigma\sqrt{2\log(2n)}$.

Proof.

$$\begin{aligned}\mathbb{E}\max_{j=1,\dots,n}|X_j| &= \mathbb{E}\max\left(\max_{j=1,\dots,n} X_j, \max_{j=1,\dots,n}(-X_j)\right) \\ &\le \mathbb{E}\max_{j=1,\dots,2n} Z_j \le \sigma\sqrt{2\log(2n)},\end{aligned}$$

where $Z_j = X_j$ when $j \le n$ and $Z_j = -X_j$ when $j > n$. □

Remark 1. The bound is sharp up to the value of the constant. Let $X_1, \dots, X_n$ be i.i.d. $N(0, \sigma^2)$. Then

$$\mathbb{E}\max_{j=1,\dots,n}|X_j| \ge c\sigma\sqrt{\log(n)}$$

for some numerical constant $c > 0$ (exercise).

Remark 2. If the random variables instead satisfy the requirement $\left(\mathbb{E}|X|_j^p\right)^{1/p} \le \sigma_p$, then

$$\mathbb{E}\max_{j=1,\dots,n}|X_j| \le \sigma_p n^{1/p}.$$

You will prove this fact as a part of your homework assignment.

Question. What if n is very large (or infinite)? The answer is given by the chaining method.

1.1 Generic chaining and Dudley's entropy integral

Remark 3. *Results in this section of the notes are presented in a different manner than we will do in class. We will follow the paper "Tail bounds via generic chaining" by S. Dirksen that can be found here: https://arxiv.org/abs/1309.3522. In particular, we will prove Theorem 3.2 in that paper.*

Let $\{X(t), t \in T\}$ be a stochastic process - a collection of random variable indexed by the set T.

Example 1. $\mathcal{F} = \{f : S \to \{\pm 1\}\}$ a collection of functions and $(X_1, Y_1), \dots (X_n, Y_n) \in S \times \{1, -1\}$ i.i.d. from P. Define

$$Z_n(f) := \frac{1}{n}\sum_{j=1}^{n} \mathbb{I}_{\{Y_j \neq f(X_j)\}} - \Pr(Y \neq f(X)).$$

$\{Z_n(f),\ f \in \mathcal{F}\}$ is called the *empirical process* indexed by the set $\mathcal{F}$ (in particular, for each $f \in \mathcal{F}$, $Z_n(f)$ is a random variable).

We want to control the size of $\mathbb{E}\sup\limits_{t\in T}(X(t) - \mathbb{E}X(t))$. To this end, define

$$\mathbb{E}\sup_{t\in T} X(t) := \sup_{S\subseteq T,\ S \text{ is finite}} \{\mathbb{E}\max_{t\in S} X(t)\}.$$

This definition allows to avoid potential measurability issues associated with the supremum. From now on, we can assume that T is finite (however, we are seeking for the bounds that do not depend on the cardinality of T explicitly).

Assume that T is equipped with a (pseudo)metric $d : T \times T \to \mathbb{R}$, so that (T, d) is a metric space. For $S \subset T, t \in T$, define

$$d(t, S) = \inf_{s\in S} d(t, s).$$

Assumption. The process $\{X(t), t \in T\}$ has sub-Gaussian increments with respect to $d(\cdot,\cdot)$, meaning that

$$\forall t_1, t_2 \in T, \quad X(t_1) - X(t_2) \in \mathrm{SG}\left(d^2(t_1, t_2)\right).$$

Example 2. Let $W(t),\ t \in [0, 1]$ be the usual Brownian motion (meaning that $W(t)$ is a centered Gaussian process such that $W(0) = 0$ and $\mathrm{Cov}(W(t), W(s)) = \min(t, s)$). Then it has sub-Gaussian increments with respect to $d(s, t) = \sqrt{|t - s|}$.

Let $T_0 \subseteq T_2 \subseteq \dots \subseteq T_k \subseteq \ldots \subseteq T$ be a sequence of subsets such that $\mathrm{card}(T_0) = 1$ and $\mathrm{card}(T_k) = 2^{2^k}$. In this case, we will say that $\{T_j\}_{j=1}^{\infty}$ is an *admissible sequence*. Cardinality growth rate of admissible sequence is motivated by the fact that $\mathrm{card}(T_k)^2 = \mathrm{card}(T_{k+1})$.

Theorem 1 (Generic Chaining Bound). Let $\{T_j\}_{j=1}^{\infty}$ be an admissible sequence such that $T_0 = \{t_0\}$. Then $\forall u > 0$,

$$\Pr\left(\sup_{t\in T}|X(t) - X(t_0)| \geq 2u \sup_{t\in T}\sum_{j=0}^{\infty} 2^{j/2} d(t, T_j)\right) \leq \frac{128}{7} e^{-u^2/4}.$$

The quantity

$$\inf_{\{T_j\}_{j\geq 0}} \sup_{t\in T} \sum_{j\geq 0} 2^{j/2} d(t, T_j),$$

where the inf is taken over all admissible sequences, is called the **generic chaining complexity** that is denoted by $\gamma_2(T,d)$. Note that it depends only on the metric space (T,d) itself.

Proof. Define $\pi_j : T \to T_j$ via

$$\pi_j(t) = \arg\min_{s\in T_j} d(t,s),$$

the closest point to t in T_j. In particular, $d(t, \pi_j t) = d(t, T_j)$. Recall that $T_0 = \{t_0\}$ and note that

$$X(t) - X(t_0) = \sum_{j=0}^{\infty} (X(\pi_{j+1}t) - X(\pi_j t)).$$

This "telescoping sum" is the main idea of generic chaining. Note that the sum is in fact finite: since we assumed that T itself is finite, $T_N = T$ for some (possibly very large) N. Next, note that

$$d(\pi_j t, \pi_{j+1} t) \leq d(t, \pi_j t) + d(t, \pi_{j+1} t) = d(t, T_j) + d(t, T_{j+1})$$

Given $u > 0$, define the "nice event"

$$\mathcal{E} = \left\{ |X(s_1) - X(s_2)| \leq u 2^{j/2} d(s_1, s_2), \forall s_1 \in T_j, s_2 \in T_{j+1}, \forall j \geq 0 \right\}.$$

Note that when $\mathcal{E}$ holds,

$$\begin{aligned}
\sup_t |X(t) - X(t_0)| &= \sup_t \left| \sum_{j=0}^{\infty} (X(\pi_{j+1}t) - X(\pi_j t)) \right| \leq \sup_t \sum_{j=0}^{\infty} |(X(\pi_{j+1}t) - X(\pi_j t))| \\
&\leq \sup_t \sum_{j=0}^{\infty} u\, 2^{j/2} \left(d(t, T_j) + d(t, T_{j+1}) \right) \leq (1 + 1/\sqrt{2}) u \sup_t \sum_{j=0}^{\infty} u\, 2^{j/2} d(t, T_j).
\end{aligned}$$

Hence, it remains to control $\Pr(\mathcal{E}^c)$. To this end, we use the union bound and the tail bound for sub-Gaussian random variables (Corollary **??**):

$$\begin{aligned}
\Pr(\mathcal{E}^c) &= \Pr\left(\bigcup_{j\geq 0, s_1\in T_j, s_2\in T_{j+1}} |X(s_1) - X(s_2)| > u 2^{j/2} d(s_1, s_2) \right) \\
&\leq \sum_{j\geq 0} \sum_{s_1\in T_j, s_2\in T_{j+1}} \Pr\left(|X(s_1) - X(s_2)| > u 2^{j/2} d(s_1, s_2) \right) \\
&\leq \sum_{j\geq 0} \sum_{s_1\in T_j, s_2\in T_{j+1}} 2 e^{\frac{-u^2 2^j d^2(s_1,s_2)}{2d^2(s_1,s_2)}} = \sum_{j\geq 0} \sum_{s_1\in T_j, s_2\in T_{j+1}} 2 e^{-u^2 2^{j-1}} \\
&\leq \sum_{j\geq 0} \operatorname{card}(T_j) \operatorname{card}(T_{j+1}) \cdot 2 e^{-u^2 2^{j-1}} = 2 \sum_{j\geq 0} 2^{2^j + 2^{j+1}} e^{-u^2 2^{j-1}} \\
&= 2 \sum_{j\geq 0} e^{-u^2 2^{j-1} + 6 2^{j-1} \log 2} = 2 \sum_{j\geq 0} e^{-2^{j-1}\left(u^2 - 6\log 2\right)}.
\end{aligned}$$

Assume that $\frac{u^2}{2} > 6\log 2$. Then

$$\begin{aligned}\Pr(\mathcal{E}^c) \leq & 2\sum_{j\geq 0}\exp\left(-2^j\frac{u^2}{4}\right) \leq 2\sum_{j\geq 0}\exp\left(-(j+1)\frac{u^2}{4}\right) \\ = & 2\frac{\exp\left(-\frac{u^2}{4}\right)}{1-\exp\left(-\frac{u^2}{4}\right)} \leq \frac{2}{7/8}\exp\left(-\frac{u^2}{4}\right).\end{aligned}$$

When $u^2/2 \leq 6\log 2$, $\frac{16}{7}\exp\left(-\frac{u^2}{4}\right) \leq \frac{16}{7}\frac{1}{8}$, hence $8\frac{16}{7}\exp\left(-\frac{u^2}{4}\right) \leq 1$, thus the inequality

$$\Pr(\mathcal{E}^c) \leq 8\frac{16}{7}\exp\left(-\frac{u^2}{4}\right)$$

holds for all $u > 0$.

$\square$

Exercise. Show that

$$\mathbb{E}\sup_{t\in T}|X(t) - X(t_0))| \leq C\gamma_2(T,d)) \tag{1.1}$$

for some numerical constant C (use the fact that for any nonnegative random variable Y, $\mathbb{E}[Y] = \int_0^\infty \mathbb{P}(Y \geq t)dt$). Moreover, demonstrate that

$$\sup_{t\in T}|X(t) - X(t_0)| - \mathbb{E}\sup_{t\in T}|X(t) - X(t_0)| \in \mathrm{SG}(C_1\gamma_2(T,d))$$

for some numerical constant C_1 (however, the value of the sub-Gaussian parameter can be improved).