MATH 281C: Mathematical Statistics

Lecture 7

The main goal for today is to state and prove Dudley's entropy bound for the suprema of sub-Gaussian processes. The proof involves an idea called chaining, which Kolmogorov pioneered. Before we start with chaining, let us recall the following elementary bound on the Rademacher average that was covered in Lecture 5.

Proposition 0.1. Suppose T is a finite subset of $\mathbb{R}^n$ with cardinality $|T|$. Then

$$R_n(T) = \mathbb{E}\max_{t\in T}\left|\frac{1}{n}\sum_{i=1}^{n}\epsilon_i t_i\right| \le \sqrt{\frac{6\log(2|T|)}{n}}\max_{t\in T}\sqrt{\frac{1}{n}\sum_{i=1}^{n}t_i^2}.$$

1 Dudley's Metric Entropy Bound

What we need later is a stronger version than the previous result.

Proposition 1.1. Let T be a finite set and let $\{X_t, t\in T\}$ be a stochastic process. Suppose that for every $t\in T$ and $u\ge 0$, the inequality

$$\mathbb{P}(|X_t|\ge u)\le 2e^{-u^2/(2\sigma^2)} \tag{7.1}$$

holds, where $\sigma > 0$ is a constant (variance proxy). Then, for a universal constant $C>0$ we have

$$\mathbb{E}\Big(\max_{t\in T}|X_t|\Big)\le C\sigma\sqrt{\log(2|T|)}. \tag{7.2}$$

Remark 1.1. Note that Proposition 1.1 is indeed a generalization of Proposition 0.1. This is because for $X_t=\sum_{i=1}^n \epsilon_i t_i$ (with $t\in T$), Hoeffding's inequality assures that (7.2) holds with

$$\sigma^2 = \max_{t\in T}\sum_{i=1}^{n}t_i^2.$$

Proposition 1.1 holds for every set of random variables X_t satisfying (7.1) so in addition to $X_t = \sum_{i=1}^n \epsilon_i t_i$, it also holds for $X_t\sim\mathcal{N}(0,\sigma^2)$.

Proof of Proposition 1.1. Because

$$\mathbb{E}\Big(\max_{t\in T}|X_t|\Big) = \int_0^\infty \mathbb{P}\Big(\max_{t\in T}|X_t|\ge u\Big)\mathrm{d}u,$$

we can control $\mathbb{E}\max_{t\in T}|X_t|$ by bound the tail inequality $\mathbb{P}(\max_{t\in T}|X_t|\ge u)$ for every $u\ge 0$. For this, write

$$\mathbb{P}\Big(\max_{t\in T}|X_t|\ge u\Big) = \mathbb{P}\Big[\bigcup_{t\in T}\{|X_t|\ge u\}\Big]\le\sum_{t\in T}P(|X_t|\ge u)\le 2|T|e^{-u^2/(2\sigma^2)}.$$

This bound is good for large u but not so good for small u (because $|T|$ is large). It is therefore good to use it only for $u \geq u_0$ for some u_0 to be specified later. This gives

$$\begin{aligned}
\mathbb{E}\Big(\max_{t\in T} |X_t|\Big) &= \int_0^\infty \mathbb{P}\Big(\max_{t\in T} |X_t \geq u\Big) \mathrm{d}u \\
&= \int_0^{u_0} \mathbb{P}\Big(\max_{t\in T} |X_t \geq u\Big) \mathrm{d}u + \int_{u_0}^\infty \mathbb{P}\Big(\max_{t\in T} |X_t \geq u\Big) \mathrm{d}u \\
&\leq u_0 + 2|T| \int_{u_0}^\infty e^{-u^2/(2\sigma^2)} \mathrm{d}u \\
&\leq u_0 + \frac{2|T|}{u_0} \sigma^2 e^{-u_0^2/(2\sigma^2)},
\end{aligned}$$

valid for any $u_0 > 0$. One can try to minimize the above term over u_0. A simpler strategy is to realize that the large term here is $2|T|$ so one can choose u_0 to kill this term by setting

$$e^{u_0^2/(2\sigma^2)} = 2|T| \quad \text{or} \quad u_0 = \sigma\sqrt{2\log(2|T|)}.$$

This gives $\mathbb{E}\max_{t\in T} |X_t| \leq \sigma\sqrt{2\log(2|T|)} + \sigma/\sqrt{2\log(2|T|)} \leq C\sigma\sqrt{\log(2|T|)}$ for some $C > \sqrt{2}$. □

The bound in (7.2) can be loose when many of the X_t's are close to each other: for instance, when $X_t, t \in T$ are highly correlated, $\max_{t\in T}|X_t| \approx X_{t_0}$ for a single $t_0 \in T$, and hence the bound in (7.2) is loose by a factor of $\log|T|$. However there exist examples where the bound in (7.2) is tight. The simplest example is the following. Suppose $X_t, t \in T$ are independently distributed as $\mathcal{N}(0, \sigma^2)$. Then it can be shown that

$$\mathbb{E}\Big(\max_{t\in T} X_t\Big) \geq c\sigma\sqrt{\log|T|}$$

for a positive constant $c > 0$. Therefore, in this case (under independence), (7.2) is tight up to constant factor. A proof of this lower bound can be found here. This example means that Proposition 1.1 cannot be improved without additional assumptions on the process $\{X_t, t \in T\}$. Chaining gives improved bounds for $\mathbb{E}\max_{t\in T} X_t$ or $\mathbb{E}\max_{t\in T}|X_t|$ under an assumption on $\{X_t, t \in T\}$ that is different from (7.1). The assumption (7.1) pertains to the marginal distribution of each X_t but does not say anything about how close X_t is to another X_s when t and s are close. In contrast, for chaining, one assumes the existence of a metric d on T such that

$$\mathbb{P}(|X_s - X_t| \geq u) \leq 2e^{-u^2/\{2d^2(s,t)\}}. \tag{7.3}$$

Under this assumption, chaining provides a bound on $\mathbb{E}\max_{t\in T}|X_t|$ which involves the metric properties of (T, d).

1.1 Dudley's bound for finite T

We first state Dudley's bound when the index set T is finite, and subsequently improve it to the case when T is infinite.

Theorem 1.1 (Dudley's metric entropy bound for finite T)**.** Suppose (T, d) is a finite metric space and $\{X_t, t \in T\}$ is a stochastic process such that the bound (7.3) holds for every $s, t \in T$ and $u \geq 0$. Then, for a universal positive constant C, the following inequality holds for every $t_0 \in T$:

$$\mathbb{E}\Big(\max_{t\in T} |X_t - X_{t_0}|\Big) \leq C\int_0^\infty \sqrt{\log M(\epsilon, T, d)}\, \mathrm{d}\epsilon. \tag{7.4}$$

The following remarks mention some alternative forms of inequality (7.4) and also describe some implications.

(i) Let D denote the diameter of the metric space T, i.e. $D = \max_{s,t\in T} d(s,t)$. Then the packing number $M(\epsilon, T, d)$ equals 1 for $\epsilon \geq D$ (it is impossible to have two points in T whose distance is strictly larger than ϵ when $\epsilon > D$). Therefore,

$$\int_0^\infty \sqrt{\log M(\epsilon, T, d)}\, d\epsilon = \int_0^D \sqrt{\log M(\epsilon, T, d)}\, d\epsilon.$$

Moreover,

$$\begin{aligned}\int_0^D \sqrt{\log M(\epsilon, T, d)}\, d\epsilon &= \int_0^{D/2} \sqrt{\log M(\epsilon, T, d)}\, d\epsilon + \int_{D/2}^D \sqrt{\log M(\epsilon, T, d)}\, d\epsilon \\ &= \int_0^{D/2} \sqrt{\log M(\epsilon, T, d)}\, d\epsilon + \int_0^{D/2} \sqrt{\log M(\epsilon + D/2, T, d)}\, d\epsilon \\ &\leq 2\int_0^{D/2} \sqrt{\log M(\epsilon, T, d)}\, d\epsilon,\end{aligned}$$

because $M(\epsilon + D/2, T, d) \leq M(\epsilon, T, d)$. We can thus state Dudley's bound as

$$\mathbb{E}\Big(\max_{t\in T} |X_t - X_{t_0}|\Big) \leq C \int_0^{D/2} \sqrt{\log M(\epsilon, T, d)}\, d\epsilon,$$

where the C above equals twice the constant C in (7.4). Similarly, again by splitting the above integral in two parts (over 0 to $D/4$ and over $D/4$ to $D/2$), we can also state Dudley's bound as

$$\mathbb{E}\Big(\max_{t\in T} |X_t - X_{t_0}|\Big) \leq C \int_0^{D/4} \sqrt{\log M(\epsilon, T, d)}\, d\epsilon.$$

The constant C above now is 4 times the constant in (7.4).

(ii) Inequality (7.4) implies

$$\mathbb{E} \max_{t\in T} |X_t| \leq \mathbb{E}|X_{t_0}| + C \int_0^{D/2} \sqrt{\log M(\epsilon, T, d)}\, d\epsilon \quad \text{for every } t_0 \in T.$$

(iii) If $X_t, t \in T$ are joint centered Gaussian, then $X_t - X_s$ is zero-mean normal random variable for every $s, t \in T$ so that (7.3) holds with

$$d(s,t) = \sqrt{\mathbb{E}(X_s - X_t)^2}.$$

(iv) The advantage of Theorem 1.1 over Proposition 1.1 is clear from the following example. Suppose $X_t, t \in T$ are given by

$$X_t = X_0 + \eta Z_t$$

for some very small $\eta > 0$ ($\eta \ll \sigma$), $X_0 \sim \mathcal{N}(0, \sigma^2)$ and $Z_t, t \in T$ are iid standard normal variables. Noting that $\mathrm{var}(X_t) = \sigma^2 + \eta^2$, Proposition 1.1 implies $\mathbb{E}\max_{t\in T} |X_t| \lesssim \sigma\sqrt{\log |T|}$. On the other hand, because

$$d(s,t) = \sqrt{\mathbb{E}(X_t - X_s)^2} \leq \eta\sqrt{2},$$

the packing number $M(\epsilon, T, d)$ equals 1 for all but sufficiently small values of ϵ, say for $\epsilon > \epsilon_0$. Thus Dudley's bound will give $\sigma + C\epsilon_0 \sqrt{\log|T|}$, which can be much smaller than the bound given by Proposition 1.1 (because ϵ_0 is small, i.e. $\epsilon_0 \ll \sigma$).

We will now give the proof of Theorem 1.1, which is based on an idea called chaining. Specifically, we will split $\max_{t\in T}(X_t - X_{t_0})$ in chains, and use the bound given by Proposition 1.1 within the links of each chain.

Proof of Theorem 1.1. Let T be the diameter of D. For $n \geq 1$, let T_n be a maximal $D2^{-n}$-separated subset of T, i.e. $\min_{s,t\in T_n, s\neq t} d(s,t) > D2^{-n}$, so that $|T_n| = M(D2^{-n}, T, d)$. Due to its maximality,

$$\max_{t\in T} \min_{s\in T_n} d(s,t) \leq D2^{-n}. \tag{7.5}$$

Because T is finite and $d(s,t) > 0$ for all $s \neq t \in T$, the set T_n will equal T when n is large. Let

$$N = \min\{n \geq 1 : T_n = T\}.$$

For each $n \geq 1$, let $\pi_n : T \to T_n$ denote the function which maps each point $t \in T$ to the point in T_n that is closest to T. If there are multiple such points, choose one arbitrarily. In other words, $\pi_n(t)$ is chosen so that

$$d(t, \pi_n(t)) = \min_{s\in T_n} d(t,s).$$

By (7.5),

$$d(t, \pi_n(t)) \leq D2^{-n} \quad \text{for all } t \in T \text{ and } n \geq 1. \tag{7.6}$$

In particular, $\pi_N(t) = t$. Moreover, let $T_0 = \{t_0\}$ so that $\pi_0(t) = t_0$ for all $t \in T$.

Using the maps $\pi_0, \pi_1, \ldots, \pi_N$ to define a telescoping sum

$$X_t - X_{t_0} = \sum_{n=1}^{N} \{X_{\pi_n(t)} - X_{\pi_{n-1}(t)}\}, \qquad \text{for every } t \in T. \tag{7.7}$$

The sequence

$$t_0 \to \pi_1(t) \to \pi_2(t) \to \cdots \to \pi_{N-1}(t) \to \pi_N(t) = t$$

can be viewed as a chain from t_0 to t. This is what gives the argument the name *chaining*.

By (7.7), we have

$$\max_{t\in T} |X_t - X_{t_0}| \leq \max_{t\in T} \sum_{n=1}^{N} |X_{\pi_n(t)} - X_{\pi_{n-1}(t)}| \leq \sum_{n=1}^{N} \max_{t\in T} |X_{\pi_n(t)} - X_{\pi_{n-1}(t)}|,$$

so that

$$\mathbb{E} \max_{t\in T} |X_t - X_{t_0}| \leq \sum_{n=1}^{N} \mathbb{E} \max_{t\in T} |X_{\pi_n(t)} - X_{\pi_{n-1}(t)}|. \tag{7.8}$$

Now to bound $\mathbb{E}\max_{t\in T}|X_{\pi_n(t)} - X_{\pi_{n-1}(t)}|$ for each $1 \le n \le N$, we will use the elementary bound given by Proposition 1.1. By (7.3),

$$\mathbb{P}\{|X_{\pi_n(t)} - X_{\pi_{n-1}(t)}| \ge u\} \le 2\exp\left\{\frac{-u^2}{2d^2(\pi_n(t),\pi_{n-1}(t))}\right\}.$$

Note that

$$d(\pi_n(t),\pi_{n-1}(t)) \le d(\pi_n(t),t) + d(\pi_{n-1}(t),t) \le D2^{-n} + D2^{-n+1} = 3D2^{-n}.$$

Thus, Proposition 1.1 can be applied with $\sigma := 3D2^{-n}$, implying

$$\mathbb{E}\max_{t\in T}|X_{\pi_n(t)} - X_{\pi_{n-1}(t)}| \le C\frac{3D}{2^n}\sqrt{\log(2|T_n||T_{n-1}|)} \le C_1\frac{D}{2^n}\sqrt{\log(2M(D2^{-n},T,d))}.$$

Plugging this into (7.8), and applying the monotonicity of $\epsilon \mapsto M(\epsilon,T,d)$, we deduce that

$$\begin{aligned}
\mathbb{E}\max_{t\in T}|X_t - X_{t_0}| &\le C_1\sum_{n=1}^{N}\frac{D}{2^n}\sqrt{\log(2M(D2^{-n},T,d))} \\
&\le 2C_1\sum_{n=1}^{N}\int_{D2^{-(n+1)}}^{D2^{-n}}\sqrt{\log(2M(\epsilon,T,d))}\,\mathrm{d}\epsilon \\
&= 2C_1\int_{D2^{-N-1}}^{D/2}\sqrt{\log(2M(\epsilon,T,d))}\,\mathrm{d}\epsilon \\
&\le 2C_1\int_{0}^{D/2}\sqrt{\log(2M(\epsilon,T,d))}\,\mathrm{d}\epsilon \\
&\le 4C_1\int_{0}^{D/4}\sqrt{\log(2M(\epsilon,T,d))}\,\mathrm{d}\epsilon.
\end{aligned}$$

For $\epsilon \le D/4$, the packing number $M(\epsilon,T,d) \ge 2$ so that $\log(2M(\epsilon,T,d)) \le 2\log M(\epsilon,T,d)$. It then follows that

$$\mathbb{E}\max_{t\in T}|X_t - X_{t_0}| \le C_2\int_0^{D/4}\sqrt{\log M(\epsilon,T,d)}\,\mathrm{d}\epsilon,$$

as desired. □

1.2 Dudley's bound for infinite T

We next prove Dudley's bound for the case of infinite T. This requires a technical assumption called *separability*, which will always be satisfied in our applications.

Definition 1.1 (Separable stochastic process)**.** . Let (T,d) be a metric space. The stochastic process $\{X_t, t\in T\}$ indexed by T is said to be separable if there exists a null set Ω_0 (of the probability space Ω) and a countable subset $\widetilde{T}$ of T such that for all $\omega \notin \Omega_0$ and $t \in T$, there exists a sequence $\{t_n\} \subseteq \widetilde{T}$ with $\lim_{n\to\infty} d(t_n,t) = 0$ and $\lim_{n\to\infty} X_{t_n}(\omega) = X_t(\omega)$.

The definition of separability requires that $\widetilde{T}$ is a dense subset of T, meaning that the metric space (T,d) is separable (a metric space is said to be separable if it has a countable dense subset).

The following fact is easy to check: if (T,d) is a separable metric space and if $X_t, t\in T$ has continuous sample paths (almost surely), then $X_t, t\in T$ is separable. The statement that $X_t, t\in T$

has continuous sample paths (almost surely) means that there exists a null set Ω_0 such that for all $\omega \in \Omega_0$, the function $t \mapsto X_t(\omega)$ is continuous on T.

We have also the following fact: if $\{X_t, t \in T\}$ is a separable stochastic process, then

$$\sup_{t \in T} |X_t - X_{t_0}| = \sup_{t \in \widetilde{T}} |X_t - X_{t_0}| \quad \text{almost surely,} \tag{7.9}$$

for every $t_0 \in T$. Here $\widetilde{T}$ is a countable subset of T, which appears in the definition of separability of $X_t, t \in T$. In particular, the statement (7.9) implies that $\sup_{t \in T} |X_t - X_{t_0}|$ is measurable (note that uncountable suprema are in general not guaranteed to be measurable; but this is not an issue for separable processes).

Next we state Dudley's theorem for separable processes. This theorem does not impose any cardinality restrictions on T (it holds for both finite and infinite T).

Theorem 1.2. Suppose (T, d) is a separable metric space and let $\{X_t, t \in T\}$ be a separable stochastic process. Suppose that (7.3) holds for every $s, t \in T$ and $u \geq 0$. Then, for every $t_0 \in T$, we have

$$\mathbb{E}\left(\max_{t \in T} |X_t - X_{t_0}|\right) \leq C \int_0^{D/2} \sqrt{\log M(\epsilon, T, d)} \, \mathrm{d}\epsilon,$$

where D is the diameter of the metric space (T, d).

Proof. Let $\widetilde{T}$ be a countable subset of T such that (7.9) holds. We may assume $\widetilde{T}$ contains t_0 (otherwise simply add t_0 to $\widetilde{T}$). Since $\widetilde{T}$ is countable, write it as $\widetilde{T} = \{t_0, t_1, t_2, \ldots\}$, and for each $k \geq 1$, let $\widetilde{T}_k$ be the finite set obtained by taking the first k elements of $\widetilde{T}$. Then $\widetilde{T}_k$ contains t_0 for every $k \geq 1$.

Applying Theorem 1.1, the finite index set version of Dudley's theorem, to $\{X_t, t \in \widetilde{T}_k\}$, we obtain

$$\mathbb{E} \max_{t \in \widetilde{T}_k} |X_t - X_{t_0}| \leq C \int_0^{\mathrm{diam}(\widetilde{T}_k)/2} \sqrt{\log M(\epsilon, \widetilde{T}_k, d)} \, \mathrm{d}\epsilon \leq C \int_0^{D/2} \sqrt{\log M(\epsilon, T, d)} \, \mathrm{d}\epsilon.$$

Note that the RHS does not depend on k. Letting $k \to \infty$ on the LHS, we use the monotone convergence theorem to obtain

$$\mathbb{E} \max_{t \in \widetilde{T}} |X_t - X_{t_0}| \leq C \int_0^{D/2} \sqrt{\log M(\epsilon, T, d)} \, \mathrm{d}\epsilon.$$

Together with (7.9), this completes the proof. □

1.3 Application of Dudley's bound to Rademacher averages

Suppose $T \subseteq \mathbb{R}^n$ and consider the stochastic process $X_t, t \in T$ given by

$$X_t = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \epsilon_i t_i,$$

where $\epsilon_1, \ldots, \epsilon_n$ are iid Rademacher random variables.

Let us define the following norm on $\mathbb{R}^n$:

$$\|t\|_n = \sqrt{\frac{1}{n}\sum_{i=1}^n t_i^2}.$$

Also, let $d_n(s,t) = \|s-t\|_n$ be the corresponding metric on $\mathbb{R}^n$.

By Hoeffding's inequality, for every $u \geq 0$,

$$\mathbb{P}(|X_t - X_s| \geq u) \leq 2\exp\left\{\frac{-nu^2}{2\sum_{i=1}^n (s_i - t_i)^2}\right\} = 2\exp\left(\frac{-u^2}{2\|s-t\|_n^2}\right) = 2\exp\left\{\frac{-u^2}{2d_n^2(s,t)}\right\}.$$

Hence $X_t, t \in T$ satisfies the assumptions in Dudley's theorems with the metric d_n. Also note that $T = \mathbb{R}^n$ is trivially separable, and that the map

$$t = (t_1, \ldots, t_n) \mapsto \frac{1}{\sqrt{n}}\sum_{i=1}^n \epsilon_i t_i$$

is linear (and hence continuous in t). This means that $X_t, t \in T$ is separable. We then apply Dudley's theorem with $t_0 = (0, \ldots, 0)$. Since t_0 may not be contained in T, we can apply Theorem 1.2 to $T \cup \{0\}$ and notice that $\sup_{t \in T \cup \{0\}} |\cdot| = \sup_{t \in T} |\cdot|$. As a result,

$$\mathbb{E}\sup_{t \in T}\left|\frac{1}{\sqrt{n}}\sum_{i=1}^n \epsilon_i t_i\right| \leq C\int_0^{\mathrm{diam}(T \cup \{0\})/2} \sqrt{\log M(\epsilon, T \cup \{0\}, d_n)}\, \mathrm{d}\epsilon,$$

where the diameter and packing numbers above are with respect to the d_n metric. It is easy to see that

$$\mathrm{diam}(T \cup \{0\}) = \sup_{s,t \in T \cup \{0\}} \|s-t\|_n \leq 2\sigma_n \quad \text{with} \quad \sigma_n := \sup_{t \in T}\|t\|_n.$$

We thus obtain the following upper bound

$$\mathbb{E}\sup_{t \in T}\left|\frac{1}{\sqrt{n}}\sum_{i=1}^n \epsilon_i t_i\right| \leq C\int_0^{\sigma_n} \sqrt{\log M(\epsilon, T \cup \{0\}, d_n)}\, \mathrm{d}\epsilon.$$

In the next class, we will combine the above bound with the *symmetrization* technique to obtain an important upper bound on the suprema of empirical processes.