ORIE 7790 High Dimensional Probability and Statistics	Lecture 11 - 02/27/2020

Lecture 11: Random Processes and Chaining

Lecturer: Yudong Chen *Scribe: Billy Jin*

Reading:

- Sec 7.4 and 8.1 of Vershynin book,
- Sec 5.3 of Wainwright.

In this lecture, we give two examples where using Sudakov's lower bound in reverse can be used to bound the covering number of a set. We then prove Dudley's entropy integral upper bound and apply it to prove a uniform law of large numbers.

1 Applications of Sudakov to Bounding Covering Number

First, recall Sudakov's minorization inequality from last lecture.

Theorem 1. *Let $(Z_\theta)_{\theta \in T}$ be a zero-mean Gaussian process. Then*

$$\mathbb{E}\left[\sup_{\theta \in T} Z_\theta\right] \gtrsim \epsilon\sqrt{\log N(\epsilon, T, \rho)},$$

where $\rho(\theta, \theta') := \sqrt{\mathbb{E}(Z_\theta - Z_{\theta'})^2}$.

1.1 Covering Number of the ℓ_1-Ball

Let $\mathbb{B}_1^d := \{\theta \in \mathbb{R}^d : \|\theta\|_1 \le 1\}$ be the ℓ_1 unit ball in $\mathbb{R}^d$. Consider the canonical Gaussian process:

$$Z_\theta = \langle \theta, g \rangle \quad \text{for all } \theta \in \mathbb{B}_1^d,$$

where $g \sim N(0, I_d)$. Recall from last lecture that the canonical metric for this process is $\rho(\theta, \theta') = \|\theta - \theta'\|_2$. Applying Sudakov's inequality, we obtain that, for all $\epsilon > 0$,

$$\begin{aligned}
\epsilon\sqrt{\log N(\epsilon, \mathbb{B}_1^d, \|\cdot\|_2)} &\lesssim \mathbb{E}\left[\sup_{\theta \in \mathbb{B}_1^d} \langle \theta, g \rangle\right] \\
&\le \mathbb{E}\left[\sup_{\theta \in \mathbb{B}_1^d} \|\theta\|_1 \|g\|_\infty\right] \\
&= \mathbb{E}\left[\, \|g\|_\infty\right]
\end{aligned}$$

The first inequality is Sudakov's bound, and the second inequality is by Hölder's inequality. (The second inequality is also easy to see directly.) Hence it suffices to bound $\mathbb{E}\|g\|_\infty$; the following lemma tells us how to do this.

Lemma 1. *Let g_i be a σ^2-sub-Gaussian RV for each $i = 1, \ldots, d$. Then $\mathbb{E}\max_i |g_i| \lesssim \sigma\sqrt{\log d}$.*

Proof We have, for all $\beta > 0$,

$$\begin{aligned}
\mathbb{E} \max |g_i| &= \frac{1}{\beta} \mathbb{E} \log e^{\beta \max |g_i|} \\
&= \frac{1}{\beta} \mathbb{E} \log e^{\beta \cdot \max\{g_i, -g_i\}} \\
&= \frac{1}{\beta} \mathbb{E} \log \max\{e^{\beta g_i}, e^{-\beta g_i}\} \\
&\leq \frac{1}{\beta} \mathbb{E} \log \left(\sum_{i=1}^{d} e^{\beta g_i} + \sum_{i=1}^{d} e^{-\beta g_i} \right) && \max \leq \text{sum} \\
&\leq \frac{1}{\beta} \log \mathbb{E} \left(\sum_{i=1}^{d} e^{\beta g_i} + \sum_{i=1}^{d} e^{-\beta g_i} \right) && \text{Jensen's} \\
&= \frac{1}{\beta} \log \left(2d \, \mathbb{E} \, e^{\beta g} \right) && \text{linearity of expectation} \\
&\leq \frac{1}{\beta} \log \left(2d \cdot e^{\beta^2 \sigma^2 / 2} \right) && \text{MGF definition of sub-Gaussian RV} \\
&\lesssim \sigma \sqrt{\log d}. && \text{pick } \beta = \sqrt{\frac{2 \log d}{\sigma^2}}
\end{aligned}$$

$\square$

Remark Note that the lemma does *not* require the g_i's to be independent. Finally, the lemma still holds if each g_i is *sub-Gaussian*.

Using Lemma 1, we get

$$\epsilon \sqrt{\log N(\epsilon, \mathbb{B}_1^d, \|\cdot\|_2)} \lesssim \mathbb{E} \, \|g\|_\infty \lesssim \sqrt{\log d}.$$

Thus, the metric entropy of the ℓ_1-ball is upper bounded by

$$\log N(\epsilon, \mathbb{B}_1^d, \|\cdot\|_2) \lesssim \frac{1}{\epsilon^2} \log d.$$

Compare this with the metric entropy of the ℓ_2-ball from Lecture 8:

$$\log N(\epsilon, \mathbb{B}_2^d, \|\cdot\|_2) \lesssim d \log \left(1 + \frac{4}{\epsilon} \right).$$

We see that in high dimensions, the metric entropy of the ℓ_2-ball is much larger than that of the ℓ_1-ball.

1.2 Covering Number of a Polytope

Suppose $P \subseteq \mathbb{R}^d$ is a polytope with m vertices, with radius bounded by 1; that is, $\max_{\theta \in P} \|\theta\|_2 \leq 1$. Let $\theta^{(1)}, \ldots, \theta^{(m)}$ be the m vertices. Then, Sudakov's inequality tells us that, for all $\epsilon > 0$,

$$\epsilon \sqrt{\log N(\epsilon, P, \|\cdot\|_2)} \lesssim \mathbb{E} \sup_{\theta \in P} \langle \theta, g \rangle = \mathbb{E} \max_{i \in [m]} \left\langle \theta^{(i)}, g \right\rangle.$$

The last equality is because the maximum of a linear function over a polytope is always attained at one of the extreme points. Note that $\left\langle \theta^{(i)}, g \right\rangle \sim N(0, \left\|\theta^{(i)}\right\|_2^2)$, and $\left\|\theta^{(i)}\right\|_2 \leq 1$. Thus, by Lemma 1 (which does not require independence), we have

$$\mathbb{E} \max_{i \in [m]} \left\langle \theta^{(i)}, g \right\rangle \lesssim \sqrt{\log m}.$$

It follows that

$$\log N(\epsilon, P, \|\cdot\|_2) \lesssim \frac{1}{\epsilon^2}\log m.$$

Note that this bound is independent of the dimension d! Compare this bound with the naive bound

$$\log N(\epsilon, P, \|\cdot\|_2) \le \log N(\epsilon, \mathbb{B}_2^d, \|\cdot\|_2) \le d\log\left(1+\frac{4}{\epsilon}\right)$$

For these bounds to be equal, we need the number of vertices m to be *exponential* in the dimension d. If m is, say, polynomial in d, then the bound $\frac{1}{\epsilon^2}\log m$ we calculated for P is much better.

2 Dudley's Upper Bound

Recall: (Sub-Gaussian increments) Let $(Z_\theta)_{\theta\in T}$ be so that $Z_\theta - Z_{\theta'}$ is sub-Gaussian with parameter $\rho(\theta,\theta')^2$, for all $\theta,\theta'\in T$. Here, ρ is a metric on T.

Theorem 2 (Dudley's entropy integral bound)**.** *Suppose that $(Z_\theta)_{\theta\in T}$ is a zero-mean process with sub-Gaussian increments with respect to the metric ρ. Then*

$$\mathbb{E}\left[\sup_{\theta\in T} Z_\theta\right] \lesssim \int_0^\infty \sqrt{\log N(\epsilon, T, \rho)}\mathrm{d}\epsilon.$$

Remark Compare this with Sudakov's lower bound, which states that $\mathbb{E}\left[\,\sup_{\theta\in T} Z_\theta\,\right] \gtrsim \epsilon\sqrt{\log N(\epsilon, T, \rho)}$ for all $\epsilon > 0$. Dudley's upper bound is the area under the graph of $\sqrt{\log N(\epsilon, T, \rho)}$, whereas Sudakov's lower bound is the largest area of a rectangle under the same graph.

The proof of Dudley's upper bound uses a technique called *chaining*, which is a multi-scale version of the ϵ-net argument. To motivate, consider bounding the expected operator norm of a random matrix X:

$$\begin{aligned}
\mathbb{E}\|X\|_{\mathrm{op}} &= \mathbb{E}\sup_{u\in\mathbb{S}^{d-1}}\|Xu\|_2 \\
&\le \mathbb{E}\sup_{u_0\in S_\epsilon}\|Xu_0\|_2 + \mathbb{E}\sup_{\|u-u_0\|\le\epsilon}\|X(u-u_0)\|_2 \\
&= \mathbb{E}\sup_{u_0\in S_\epsilon}\|Xu_0\|_2 + \epsilon\,\mathbb{E}\sup_{\|u-u_0\|\le 1}\|X(u-u_0)\|_2\,.
\end{aligned}$$

The first term can be bounded by a union bound over the ϵ-net S_ϵ. Note that the second term happens to be a scaled version of what we wanted to bound. However, this is a coincidence that may not happen in general. Chaining is the technique of continuing the ϵ-net argument on the residual second term.
Proof First, a few definitions:

- Let $D := \sup_{\theta,\theta'\in T}\rho(\theta,\theta')$ be the diameter of T.
- Define the dyadic scale: $\epsilon_k := D\cdot 2^{-k}$ for $k = 0,1,2,\ldots$
- Let T_k be the smallest ϵ_k-net of T. Then $|T_k| = N(\epsilon_k, T, \rho)$.
- For $\theta\in T$, let $\pi_k(\theta)$ be the closest point in T_k to θ. So $\rho(\pi_k(\theta),\theta)\le\epsilon_k$.

Note that $T_0 = \{\theta_0\}$ for some $\theta_0\in T$, and $\pi_0(\theta) = \theta_0$ for all $\theta\in T$. Also, since the Z_θ's are zero-mean, we have

$$\mathbb{E}\sup_{\theta\in T} Z_\theta = \mathbb{E}\sup_{\theta\in T}\left(Z_\theta - Z_{\theta_0}\right).$$

To bound the RHS of the last equation, we write $Z_\theta - Z_{\theta_0}$ as a telescoping sum:

$$\begin{aligned} Z_\theta - Z_{\theta_0} &= \left(Z_{\pi_1(\theta)} - Z_{\pi_0(\theta)}\right) + \left(Z_{\pi_2(\theta)} - Z_{\pi_1(\theta)}\right) + \cdots + \left(Z_\theta - Z_{\pi_M(\theta)}\right) \\ &= \sum_{k=1}^{M} \left(Z_{\pi_k(\theta)} - Z_{\pi_{k-1}(\theta)}\right) + \left(Z_\theta - Z_{\pi_M(\theta)}\right), \end{aligned}$$

where M is any positive integer. It follows that

$$\mathbb{E} \sup_{\theta \in T} (Z_\theta - Z_{\theta_0}) \le \mathbb{E} \sum_{k=1}^{M} \sup_{\theta \in T} \left(Z_{\pi_k(\theta)} - Z_{\pi_{k-1}(\theta)}\right) + \mathbb{E} \sup_{\theta \in T} \left(Z_\theta - Z_{\pi_M(\theta)}\right).$$

Consider the kth term in the sum:

$$\mathbb{E} \sup_{\theta \in T} \left(Z_{\pi_k(\theta)} - Z_{\pi_{k-1}(\theta)}\right).$$

Recall that the RV $Z_{\pi_k(\theta)} - Z_{\pi_{k-1}(\theta)}$ is sub-Gaussian with a parameter satisfying

$$\begin{aligned} \left\| Z_{\pi_k(\theta)} - Z_{\pi_{k-1}(\theta)} \right\|_{\psi_2} &= \rho(\pi_k(\theta), \pi_{k-1}(\theta)) \\ &\le \rho(\pi_k(\theta), \theta) + \rho(\pi_{k-1}(\theta), \theta) && \text{triangle inequality of the metric } \rho \\ &\le \epsilon_k + \epsilon_{k-1} && \text{by construction} \\ &\le 2\epsilon_{k-1}. && \text{by construction} \end{aligned}$$

Thus, we have a supremum of at most $|T_k| \times |T_{k-1}|$ sub-Gaussian random variables with parameter $4\epsilon_{k-1}^2$. Using the bound on the maximum of sub-Gaussian random variables (Lemma 1), we obtain that

$$\begin{aligned} \mathbb{E} \sup_{\theta \in T} \left(Z_{\pi_k(\theta)} - Z_{\pi_{k-1}(\theta)}\right) &\lesssim \epsilon_{k-1} \sqrt{\log |T_k|\,|T_{k-1}|} \\ &\lesssim \epsilon_{k-1} \sqrt{\log |T_k|}. \end{aligned}$$

It follows that

$$\begin{aligned} \mathbb{E} \sup_{\theta \in T} (Z_\theta - Z_{\theta_0}) &\lesssim \sum_{k=1}^{M} \epsilon_{k-1} \sqrt{\log N(\epsilon_k, T, \rho)} + \mathbb{E} \sup_{\theta \in T} \left(Z_\theta - Z_{\pi_M(\theta)}\right) \\ &= \sum_{k=1}^{M} D \cdot 2^{-(k-1)} \sqrt{\log N(D \cdot 2^{-k}, T, \rho)} + \mathbb{E} \sup_{\theta \in T} \left(Z_\theta - Z_{\pi_M(\theta)}\right) \\ &\lesssim \int_{D \cdot 2^{-(M-1)}}^{D} \sqrt{\log N(\epsilon, T, \rho)} \mathrm{d}\epsilon + \mathbb{E} \sup_{\theta \in T} \left(Z_\theta - Z_{\pi_M(\theta)}\right) \\ &\le \int_0^D \sqrt{\log N(\epsilon, T, \rho)} \mathrm{d}\epsilon, \end{aligned}$$

where the last inequality is because the second term goes to zero as $M \to \infty$. (This requires a separability assumption on T; this was omitted in the lecture and is omitted in these notes.) $\square$

3 Application: Uniform Law of Large Numbers

Let $X_1, \ldots, X_n$ be iid random variables taking values in $[0, 1]$. For a fixed function $f : [0, 1] \to \mathbb{R}$, the usual Law of Large Numbers says

$$\frac{1}{n} \sum_{i=1}^{n} f(X_i) \to \mathbb{E} f(X_1) \quad \text{as } n \to \infty,$$

where the convergence is in probability or almost sure.

Can we prove *uniform* convergence over a class of functions $\mathcal{F}$? The following theorem gives one such result.

Theorem 3. *Let $\mathcal{F}$ be the set of all functions from $[0,1]$ to $\mathbb{R}$ that are 1-Lipschitz. Then*

$$\mathbb{E}\sup_{f\in\mathcal{F}}\left|\frac{1}{n}\sum_{i=1}^{n} f(X_i) - \mathbb{E}\, f(X_1)\right| \lesssim \frac{1}{\sqrt{n}}$$

Proof For any $f \in \mathcal{F}$, because f is 1-Lipschitz, we have

$$\left|\sup_{x\in[0,1]} f(x) - \inf_{x\in[0,1]} f(x)\right| \leq 1.$$

Thus, by translating if necessary, we may assume that $f : [0,1] \to [0,1]$. Consider the following empirical process $(Z_f)_{f\in\mathcal{F}}$ indexed by $f \in \mathcal{F}$:

$$Z_f := \frac{1}{n}\sum_{i=1}^{n} f(X_i) - \mathbb{E}\, f(X_1)$$

Then $\mathbb{E}\, Z_f = 0$, since the X_i's are iid. Moreover, we have

$$Z_f - Z_g = \frac{1}{n}\sum_{i=1}^{n} (f-g)(X_i) - \mathbb{E}(f-g)(X_1).$$

It follows that

$$\begin{aligned}
\|Z_f - Z_g\|_{\psi_2} &\lesssim \frac{1}{n}\left\|\sum_{i=1}^{n}(f-g)(X_i)\right\|_{\psi_2} && \text{(centering)}\\
&\lesssim \frac{1}{n}\sqrt{\sum_{i=1}^{n}\|(f-g)(X_i)\|_{\psi_2}^2} && \text{(sum of sub-Gaussians is sub-Gaussian by Hoeffding)}\\
&\lesssim \frac{1}{n}\sqrt{\sum_{i=1}^{n}\|(f-g)(X_i)\|_{\infty}^2} && \text{(bounded RVs are sub-Gaussian)}\\
&= \frac{1}{\sqrt{n}}\|f-g\|_{\infty}.
\end{aligned}$$

So, $(Z_f)_{f\in\mathcal{F}}$ has sub-Gaussian increments with respect to the metric $\rho(f,g) := \frac{1}{\sqrt{n}}\|f-g\|_\infty$. Now, applying Dudley's bound, we obtain

$$\mathbb{E}\sup_{f\in\mathcal{F}}|Z_f| \lesssim \frac{1}{\sqrt{n}}\int_0^1 \sqrt{\log N(\epsilon, \mathcal{F}, \|\cdot\|_\infty)}\,\mathrm{d}\epsilon \qquad \text{(since diameter}(\mathcal{F}) \leq 1) \tag{1}$$

Remark Note that this is not a direct application of Dudley's bound, since Dudley's inequality bounds $\mathbb{E}\sup_{f\in\mathcal{F}} Z_f$, not $\mathbb{E}\sup_{f\in\mathcal{F}}|Z_f|$. However, if we examine the proof of Dudley's inequality carefully, it actually shows that

$$\mathbb{E}\sup_{\theta\in T}|Z_\theta - Z_{\theta_0}| \lesssim \int_0^\infty \sqrt{\log N(\epsilon, T, \rho)}\,\mathrm{d}\epsilon$$

for any $\theta_0 \in T$. Taking $\theta_0 = 0 \in \mathcal{F}$ to be the zero function, we get that $\mathbb{E}\sup_{f\in\mathcal{F}}|Z_f| = \mathbb{E}\sup_{f\in\mathcal{F}}|Z_f - Z_0|$, and this is how Dudley's inequality gives us the bound in (1).

It remains to bound $N(\epsilon, \mathcal{F}, \|\cdot\|_\infty)$. To do this, we construct an *exterior* ϵ-net $\mathcal{F}_\epsilon$ of $\mathcal{F}$. (i.e., We don't require that $\mathcal{F}_\epsilon \subset \mathcal{F}$.) The construction of a usual ϵ-net is left to the homework. The construction of $\mathcal{F}_\epsilon$ is a mesh argument that covers $\mathcal{F}$ using step functions, and looks pictorially like this:

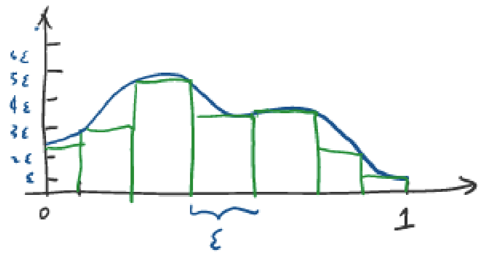

Figure 1: Covering Lipschitz functions using step functions.

One can show that $|\mathcal{F}_\epsilon| \le \left(\frac{1}{\epsilon}\right)^{\frac{1}{\epsilon}}$. (A smaller ϵ-net can be constructed; see the homework.) Plugging this into the integral in Dudley's bound, we obtain that

$$\mathbb{E} \sup_{f \in \mathcal{F}} |Z_f| \lesssim \frac{1}{\sqrt{n}} \int_0^1 \sqrt{\log \left(\frac{1}{\epsilon}\right)^{\frac{1}{\epsilon}}} \, \mathrm{d}\epsilon = \frac{\sqrt{2\pi}}{\sqrt{n}},$$

which completes the proof. $\square$

Remark Let μ be the distribution of X_i, and let μ_n be the empirical distribution:

$$\mu_n := \frac{1}{n} \sum_{i=1}^{n} \mathbb{1}_{x_i}.$$

With this notation, we have

$$\mathbb{E} \sup_{f \in \mathcal{F}} \left| \frac{1}{n} \sum_{i=1}^{n} f(X_i) - \mathbb{E} f(X_1) \right| = \mathbb{E} \sup_{f \in \mathcal{F}} \left| \int f \mathrm{d}\mu_n - \int f \mathrm{d}\mu \right|,$$

which is the *Wasserstein distance* between μ_n and μ. (The definition is equivalent to the one using transportation cost, by *Kantorovich-Rubinstein duality*).