Statistical OT Lecture 8: Chaining and Covering

1 Chaining

In the last case, we stated Dudley's chaining bound, and worked through some preliminaries.

Theorem 1. Suppose that $\mathcal{F}$ is a collection of functions, with $\|f\|_\infty \leq R$, then:

$$\mathbb{E} \sup_{f \in \mathcal{F}} \frac{1}{n} \sum_{i=1}^n f(X_i) - \mathbb{E}f \lesssim \inf_{\tau > 0} \left[\tau + \frac{1}{\sqrt{n}} \int_\tau^{2R} \sqrt{\log N(\varepsilon, \mathcal{F}, \|\cdot\|_\infty)} d\varepsilon \right].$$

1.1 Preliminaries

It will be useful to first define a sub-Gaussian process:

Definition 1. A collection of random variables $\{X_t\}_{t \in T}$ is a sub-Gaussian process if $\mathbb{E}[X_t] = 0$, and

$$\mathbb{E} \exp(\lambda(X_s - X_t)) \leq \exp(\lambda^2 \rho(s,t)^2/2), \quad \forall\ \lambda \geq 0, s, t \in T.$$

Throughout we will D denote the diameter, i.e. $D = \sup_{s,t \in T} \rho(s,t)$. Our second definition is that of a Lipschitz process:

Definition 2. A collection of random variables $\{X_t\}_{t \in T}$ is a Lipschitz process with respect to a metric ρ if for some $v > 0$,

$$|X_s - X_t| \leq v\rho(s,t) \quad \forall\ s, t \in T.$$

We were interested in a quantity of the form:

$$\mathbb{E} \sup_{f \in \mathcal{F}} \frac{1}{n} \sum_{i=1}^n f(X_i) - \mathbb{E}f,$$

and for each $f \in \mathcal{F}$ defining the random variable $X_f = \frac{1}{n}\sum_{i=1}^n f(X_i) - \mathbb{E}f$. We see that this collection of random variables has mean 0, and furthermore:

$$\begin{aligned} \mathbb{E} \exp(\lambda(X_f - X_g)) &= \prod_{i=1}^n \mathbb{E} \exp(\lambda/n(f(X_i) - g(X_i) - \mathbb{E}f(X_i) + \mathbb{E}g(X_i))) \\ &\leq \prod_{i=1}^n \exp\left(\frac{\lambda^2 \|f-g\|_\infty^2}{2n^2}\right) = \exp\left(\frac{\lambda^2 \|f-g\|_\infty^2}{2n}\right), \end{aligned}$$

where the inequality follows from Hoeffding's lemma. This shows that the empirical process we are interested in studying is sub-Gaussian with respect to the metric $\rho(s,t) = \|s-t\|_\infty/\sqrt{n}$. Furthermore, it is easy to see from the above proof that:

$$|X_f - X_g| \leq 2\|f - g\|_\infty,$$

which shows that it is also a Lipschitz process with respect to the metric $\rho(s,t) = \|s-t\|_\infty/\sqrt{n}$, with $v = \sqrt{n}$ (it will be convenient to use the same metric, and absorb the $\sqrt{n}$ scaling into v). Finally, the diameter of our collection of random variables as measured by $\rho(s,t) = \|s-t\|_\infty/\sqrt{n}$ is $\sqrt{d/n}$.

For some intuition, one should view definition 1 as capturing the "in-probability" Lipschitzness of the process, i.e. directly from the definition (and an application of Markov's inequality) we see that for a sub-Gaussian process:

$$\mathbb{P}(|X_s - X_t| \geq t\rho(s,t)) \leq \exp(-t^2/2).$$

On the other hand, definition 2 captures the "almost-sure" Lipschitzness of the process. What our example above shows (and this is typical) is that many processes of interest exhibit much stronger "in-probability" Lipschitzness versus "almost-sure" Lipschitzness. In the former case the Lipschitz constant is $O(1)$, and in the latter case it is $O(\sqrt{n})$ when the metric is chosen to be $\|s-t\|_\infty/\sqrt{n}$). We might obtain better bounds if we can correctly exploit this fact.

1.2 Proof

Now, let us finally prove Theorem 1. Let us start with a fixed X_0 which is a valid 2^{-k_0}-net for some integer k_0 (might be negative), i.e. k_0 is the largest integer such that $2^{-k_0} \geq D$. Now, we construct a sequence of ε-nets. Let N_k be a 2^{-k} net and $|N_k| = N(T, \rho, 2^{-k})$. Let $\pi_k(t)$ denote the closest point to X_t in N_k. Now, we can write for any integer m:

$$\begin{aligned}\mathbb{E}\sup_t X_t &= \mathbb{E}\sup_t (X_t - X_0) \\ &\leq \mathbb{E}\sup_t (X_t - X_{\pi_n(t)}) + \sum_{k=k_0+1}^{m} \mathbb{E}\sup_t (X_{\pi_k(t)} - X_{\pi_{k-1}(t)}).\end{aligned}$$

The first term we again control using Lipschitzness, and the second term we control using sub-Gaussianity. Note however that,

$$\rho(\pi_k(t), \pi_{k-1}(t)) \leq \rho(\pi_k(t), t) + \rho(\pi_{k-1}(t), t) \leq 3 \times 2^{-k},$$

so the k-th term in the second sum is a collection of 3×2^{-k} sub-Gaussian random variables. The k-th term is a maximum of at most $|N_k| \times |N_{k-1}| \leq |N_k|^2$ random variables. Using Lemma **??** we get that:

$$\mathbb{E}\sup_t X_t \lesssim v2^{-m} + \sum_{k=k_0+1}^{m} 2^{-k}\sqrt{\log |N_k|}.$$

Now, we note that:

$$\begin{aligned}\sum_{k=k_0+1}^{m} 2^{-k}\sqrt{\log |N_k|} &= 2\sum_{k=k_0+1}^{m} \int_{2^{-k-1}}^{2^{-k}} \sqrt{\log N(T,\rho,2^{-k})}d\varepsilon \\ &\leq 2\sum_{k=k_0+1}^{m} \int_{2^{-k-1}}^{2^{-k}} \sqrt{\log N(T,\rho,\varepsilon)}d\varepsilon \\ &\leq 2\int_{2^{-m}}^{2^{-(k_0+1)}} \sqrt{\log N(T,\rho,\varepsilon)}d\varepsilon \\ &\leq 2\int_{2^{-m}}^{D} \sqrt{\log N(T,\rho,\varepsilon)}d\varepsilon.\end{aligned}$$

Now, let us fix any $0 < \tau \leq D$, and select m to be the smallest integer such that $2^{-m} \leq \tau$, then we get:

$$\mathbb{E}\sup_t X_t \lesssim v\tau + \int_{\tau}^{D} \sqrt{\log N(T,\rho,\varepsilon)}d\varepsilon.$$

This bound applies in general. To recover the specific bound we needed for bounded functions, we simply need to use the result with $\rho(s,t) = \|f-g\|_\infty/\sqrt{n}$, $D = 2\|f\|_\infty/\sqrt{n}$, and $v = \sqrt{n}$ to obtain that:

$$\mathbb{E}\sup_{f\in\mathcal{F}} \frac{1}{n}\sum_{i=1}^{n} f(X_i) - \mathbb{E}f \lesssim \inf_{\tau>0} \sqrt{n}\tau + \int_{\tau}^{2R/\sqrt{n}} \sqrt{\log N(\mathcal{F},\rho,\varepsilon)}d\varepsilon.$$

We then note that $N(\mathcal{F}, \rho, \varepsilon) = N(\mathcal{F}, \|\cdot\|_\infty, \sqrt{n}\varepsilon)$, so via a change of variables we obtain that:

$$\mathbb{E}\sup_{f\in\mathcal{F}} \frac{1}{n}\sum_{i=1}^{n} f(X_i) - \mathbb{E}f \lesssim \inf_{\tau>0} \sqrt{n}\tau + \frac{1}{\sqrt{n}}\int_{\sqrt{n}\tau}^{2R} \sqrt{N(\mathcal{F}, \|\cdot\|_\infty, \varepsilon)}d\varepsilon,$$

which is precisely the theorem we set out to prove.

2 Covering

In the last class we used the covering numbers for the class of 1-Lipschitz functions. We note that we are interested in Lipschitz functions on $[0,1]^d$, and the quantity of interest is invariant to shifting each of the functions by a constant so without loss of generality we can focus on 1-Lipschitz functions on $[0,1]^d$ for which $f(0,\ldots,0) = 0$. Denote this class of functions as $\widetilde{\mathcal{F}}$.

Lemma 1.

$$\log N(\varepsilon, \widetilde{\mathcal{F}}, \|\cdot\|_\infty) \lesssim (4\sqrt{d}/\varepsilon)^d.$$

To prove Lemma 1 we will construct an explicit covering of the space of Lipschitz functions. I will sketch the details here. We fix an integer $j \geq 0$, $\delta > 0$ (which we will set later), and let $\mathcal{Q}_j$ denote a dyadic partition of the unit cube into cubes of side-length 2^{-j}.

Each element of $\mathcal{Q}_j$ is a cube of the form $2^{-j}([k_1, k_1+1] \times [k_2, k_2+1] \times \ldots \times [k_d, k_d+1])$, where $k_1, \ldots, k_d \in \{0, \ldots, 2^j - 1\}$. We denote these by Q_k (for a vector k). To be more rigorous, we should remove the overlaps between the cubes but we'll ignore this.

Now, consider the following class of functions $\mathcal{H}$ (they are simply piecewise, discretized, constant on the cubes defined above). $\mathcal{H}$ is all functions h which satisfy:

1. $h(x) = h_k$ for all $x \in Q_k$.
2. h_k is an integer multiple of δ.
3. $h_{(0,\ldots,0)} = 0$.
4. If $\|k - k'\|_\infty \leq 1$ (i.e. they are adjacent cubes), then $|h_k - h_{k'}| \leq 2^{-j}\sqrt{d} + \delta$.

We will first show that this collection of functions covers the set of Lipschitz functions if $2^{-j}\sqrt{d} + \delta \leq \varepsilon$, and then show that this collection is not too large. For any given Lipschitz function f, we will define a function h_f, with $(h_f)_k = \delta\lfloor f(2^{-j}(k_1, \ldots, k_d))/\delta \rfloor$ for every k. Let us check this is a function in our collection (the first 3 properties are obviously satisfied):

$$\begin{aligned}
|(h_f)_k - (h_f)_{k'}| &\leq \delta|\lfloor f(2^{-j}(k_1,\ldots,k_d))/\delta\rfloor - \lfloor f(2^{-j}(k'_1,\ldots,k'_d))/\delta\rfloor| \\
&\leq |f(2^{-j}(k_1,\ldots,k_d)) - f(2^{-j}(k'_1,\ldots,k'_d))| + \delta \\
&\leq 2^{-j}\|k - k'\|_2 + \delta \\
&\leq 2^{-j}\sqrt{d} + \delta.
\end{aligned}$$

Now we finally verify that it is a valid cover. For any $x \in Q_k$:

$$\begin{aligned}
|f(x) - h_f(x)| &\leq |f(x) - \delta\lfloor f(2^{-j}(k_1,\ldots,k_d))/\delta\rfloor| \\
&\leq |f(x) - f(2^{-j}(k_1,\ldots,k_d))| + \delta \\
&\leq \mathrm{diam}(Q_k) + \delta = 2^{-j}\sqrt{d} + \delta.
\end{aligned}$$

We will choose $\delta = 2^{-j}\sqrt{d}$. We then need to ensure that $2^{-j}\sqrt{d} \leq \varepsilon/2$, so we take j large enough for this, and we note that $2^j \leq 4\sqrt{d}/\varepsilon$ for our chosen value of j.

Now, finally, lets bound the cardinality of our collection of functions. There are 2^{dj} cubes. Now, if I fix the value of the function on a cube to some value h_k, then the value on an adjacent cube can take at most 5 distinct values. By Lipschitzness the value on an adjacent cube can differ by at most 2δ, and it has to be an integer multiple of δ. So it can take on only the values $\{h_k - 2\delta, h_k - \delta, h_k, h_k + \delta, h_k + 2\delta\}$. This means that the number of functions is at most $5^{2^{dj}}$, and we get that:

$$\log |\mathcal{H}| \lesssim 2^{dj} \lesssim (4\sqrt{d}/\varepsilon)^d.$$