Theoretical Statistics. Lecture 13.

Peter Bartlett

Metric entropy.

1. Covering number bound
2. Chaining

Recall: Covering and packing numbers

Definition: An ϵ-cover of a subset T of a pseudometric space (S, d) is a set $\hat{T} \subset T$ such that for each $t \in T$ there is a $\hat{t} \in \hat{T}$ such that $d(t, \hat{t}) \leq \epsilon$. The ϵ-covering number of T is

$$N(\epsilon, T, d) = \min\{|\hat{T}| : \hat{T} \text{ is an } \epsilon\text{-cover of } T\}.$$

An ϵ-packing of T is a subset $\hat{T} \subset T$ such that each pair $s, t \in \hat{T}$ satisfies $d(s, t) > \epsilon$. The ϵ-packing number of T is

$$M(\epsilon, T, d) = \max\{|\hat{T}| : \hat{T} \text{ is an } \epsilon\text{-packing of } T\}.$$

Recall: Covering and packing numbers

Theorem: For all $\epsilon > 0$, $M(2\epsilon) \leq N(\epsilon) \leq M(\epsilon)$.

Theorem: Let $\|\cdot\|$ be a norm on $\mathbb{R}^d$ and let B be the unit ball. Then

$$\frac{1}{\epsilon^d} \leq N(\epsilon, B, \|\cdot\|) \leq \left(\frac{2}{\epsilon} + 1\right)^d.$$

Example: If F is parameterized in a Lipschitz-continuous way by parameters in (a compact subset of) $\mathbb{R}^p$, then $N(\epsilon, F) = O(1/\epsilon^p)$.

Recall: Canonical Rademacher and Gaussian Processes

Definition: Fix a set $T \subset \mathbb{R}^n$.

1. The **canonical Gaussian process** is the stochastic process

$$G_\theta = \langle g, \theta \rangle = \sum_{i=1}^{n} g_i \theta_i,$$

   where $g_i \sim N(0,1)$ i.i.d.

2. The **canonical Rademacher process** is the stochastic process

$$R_\theta = \langle \epsilon, \theta \rangle = \sum_{i=1}^{n} \epsilon_i \theta_i,$$

   where the ϵ_i are i.i.d. and uniform on $\{\pm 1\}$.

Recall: Canonical Rademacher and Gaussian Processes

Definition: A stochastic process $\theta \mapsto X_\theta$ with indexing set T is sub-Gaussian with respect to a metric d on T if, for all $\theta, \theta' \in T$ and all $\lambda \in \mathbb{R}$,

$$\mathbf{E} \exp\left(\lambda(X_\theta - X_{\theta'})\right) \leq \exp\left(\frac{\lambda^2 d(\theta, \theta')^2}{2}\right).$$

The canonical Rademacher and Gaussian processes are sub-Gaussian wrt the Euclidean metric.

Lemma: [Finite Classes] For X_θ sub-Gaussian wrt d on T, and A a set of pairs from T,

$$\mathbf{E} \max_{(\theta,\theta')\in A} (X_\theta - X_{\theta'}) \le \max_{(\theta,\theta')\in A} d(\theta,\theta')\sqrt{2\log|A|}.$$

Covering number bound

Here's a crude approach to bounding the supremum of a sub-Gaussian process using a covering at a single scale:

Theorem: Consider a zero-mean process X_θ that is sub-Gaussian wrt the metric d on T. Suppose that the diameter of T is $D = \sup_{\theta,\theta'} d(\theta, \theta')$. Then for any ϵ,

$$\mathbf{E} \sup_{\theta} X_\theta \leq 2\mathbf{E} \sup_{d(\theta,\theta')\leq\epsilon} (X_\theta - X_{\theta'}) + 2D\sqrt{\log N(\epsilon, T, d)}.$$

Covering number bound: Proof

$$\mathbf{E}\sup_\theta X_\theta = \mathbf{E}\sup_\theta (X_\theta - X_{\theta'}) \leq \mathbf{E}\sup_{\theta,\theta'}(X_\theta - X_{\theta'}).$$

Also, if we choose $\hat{\theta} \in \hat{T}$ (a minimal ϵ-cover) with $d(\hat{\theta}, \theta) \leq \epsilon$ (and similarly for θ'), we have

$$\begin{aligned} X_\theta - X_{\theta'} &= X_\theta - X_{\hat{\theta}} + X_{\hat{\theta}} - X_{\hat{\theta}'} + X_{\hat{\theta}'} - X_{\theta'} \\ &\leq 2 \sup_{d(\theta,\hat{\theta})\leq\epsilon} (X_\theta - X_{\hat{\theta}}) + \sup_{\hat{\theta},\hat{\theta}'\in\hat{T}} X_{\hat{\theta}} - X_{\hat{\theta}'}. \end{aligned}$$

Finally, since any pair $X_\theta - X_{\theta'}$ is sub-Gaussian with parameter D^2, the Finite Lemma shows that

$$\mathbf{E} \sup_{\hat{\theta},\hat{\theta}'\in\hat{T}} X_{\hat{\theta}} - X_{\hat{\theta}'} \leq \sqrt{2D^2 \log |\hat{T}|^2} = 2D\sqrt{\log N(\epsilon, T, d)}.$$

Application: Canonical Gaussian/Rademacher process

Consider the canonical Gaussian process, $X_\theta = \langle g, \theta \rangle$ for $\theta \in T \subset \mathbb{R}^n$. Then X_θ is sub-Gaussian wrt the Euclidean metric on T. So we have

$$\mathbf{E} \sup_{d(\theta,\theta') \le \epsilon} (X_\theta - X_{\theta'}) = 2\mathbf{E} \sup_{\|v\|_2 \le \epsilon} \langle g, v \rangle \le 2\epsilon \mathbf{E}\|g\|_2 = 2\epsilon\sqrt{n}.$$

(The same argument holds for the canonical Rademacher process.) And so

$$\mathbf{E} \sup_\theta X_\theta \le 2\epsilon\sqrt{n} + 2D\sqrt{\log N(\epsilon, T, \|\cdot\|_2)}$$

Example: Canonical Gaussian process on a subspace

Consider the canonical Gaussian process with T the unit ball in a d-dimensional subspace of $\mathbb{R}^n$:
$D = 2$; $\log N(\epsilon, B, \|\cdot\|_2) \leq d \log(1 + 2/\epsilon)$.
Hence, choosing $\epsilon = \sqrt{d/n}$ gives

$$\mathbf{E} \sup_\theta X_\theta \leq 2\sqrt{d} + 4\sqrt{d \log\left(1 + 2\sqrt{n/d}\right)} = O\left(\sqrt{d \log(n/d)}\right).$$

(This is loose: the log factor is unnecessary.)

Example: Smoothly parameterized class

Suppose that F is a parameterized class, $F = \{f(\theta, \cdot) : \theta \in \Theta\}$, where $\Theta = B_2 \subset \mathbb{R}^p$. The parameterization is L-Lipschitz wrt Euclidean distance on Θ, so that for all x,

$$|f(\theta, x) - f(\theta', x)| \leq L\|\theta - \theta'\|_2.$$

Suppose also that $F = -F$ (that is, F is closed under negations).

Theorem:

$$\mathbf{E}\|R_n\|_F = O\left(L\sqrt{\frac{p\log(Ln)}{n}}\right).$$

NB: $O(\sqrt{p/n})$, plus log factor. The log factor is unnecessary.

Smoothly parameterized class: Proof

The Lipschitz condition implies that the Euclidean distance between vectors $f(\theta, X_1^n)$ is $(L\sqrt{n})$-Lipschitz wrt the Euclidean distance on Θ:

$$\sum_{i=1}^{n} |f(\theta, X_i) - f(\theta', X_i)|^2 \leq nL^2 \|\theta - \theta'\|_2^2.$$

First, exploit the fact that

$$n\mathbf{E}\|R_n\|_F = \mathbf{E} \sup_{F \cup -F} \langle \epsilon, \cdot \rangle = \mathbf{E} \sup_{F} \langle \epsilon, \cdot \rangle = \mathbf{E} \sup_{\theta} \langle \epsilon, f(\theta, X_1^n) \rangle.$$

Smoothly parameterized class: Proof

Since the process $f(\theta, X_1^n) \mapsto \langle \epsilon, f(\theta, X_1^n) \rangle$ is sub-Gaussian wrt the Euclidean norm on the vectors $f(\theta, X_1^n)$, we have

$$n\mathbf{E}\|R_n\|_F \leq 2\epsilon\sqrt{n} + \mathbf{E}4L\sqrt{n \log N(\epsilon, f(\Theta, X_1^n), \|\cdot\|_2)},$$

because $D = 2L\sqrt{n}$. Because of the Lipschitz condition,

$$N(\epsilon, f(\Theta, X_1^n), \|\cdot\|_2) \leq N(\epsilon/(L\sqrt{n}), \Theta, \|\cdot\|_2) \leq (1 + 2L\sqrt{n}/\epsilon)^p.$$

Smoothly parameterized class: Proof

Substituting $\epsilon = 1$ gives

$$\begin{aligned}\mathbf{E}\|R_n\|_F &\le \frac{2}{\sqrt{n}} + 4L\sqrt{\frac{p}{n}\log(1+2L\sqrt{n})} \\ &= O\left(L\sqrt{\frac{p\log(Ln)}{n}}\right).\end{aligned}$$

Nonparametric example: Lipschitz functions

Theorem: For F_d the set of L-Lipschitz functions (wrt $\|\cdot\|_\infty$) from $[0,1]^d$ to $[-1,1]$, there is a universal constant c_d, which depends only on d, such that

$$\mathbf{E}\|R_n\|_{F_d} \leq c_d \left(\frac{L}{n}\right)^{\frac{1}{d+2}}.$$

NB: $O(n^{-1/(d+2)})$. Even for $d=1$, this is $n^{-1/3}$, so slower than parametric. And the rate gets worse as d increases.

Nonparametric example: Proof

As before, we consider the process $f(X_1^n) \mapsto \langle \epsilon, f(X_1^n) \rangle$ for $f \in F_d$. Notice that $F_d = -F_d$. Also, the diameter of the indexing set in the Euclidean norm is $2\sqrt{n}$ (because functions in F_d can differ by at most 2). So we have

$$n\mathbf{E}\|R_n\|_F \le 2\epsilon\sqrt{n} + 4\mathbf{E}\sqrt{n \log N(\epsilon, F_d(X_1^n), \|\cdot\|_2)}.$$

Because

$$\|f(X_1^n) - f'(X_1^n)\|_2 \le \sqrt{n} \max_i |f(X_i) - f'(X_i)| \le \sqrt{n}\|f - f'\|_\infty,$$

we have $\log N(\epsilon, F_d(X_1^n), \|\cdot\|_2) \le \log N(\epsilon/\sqrt{n}, F_d, \|\cdot\|_\infty)$. Recall that $\log N(\epsilon, F_d, \|\cdot\|_\infty) = O\left((L/\epsilon)^d\right)$, so we have $\log N(\epsilon, F_d(X_1^n), \|\cdot\|_2) = O\left((L\sqrt{n}/\epsilon)^d\right)$.

Nonparametric example: Proof

Thus there is a constant c such that for sufficiently small ϵ,

$$\mathbf{E}\|R_n\|_F \leq \frac{2\epsilon}{\sqrt{n}} + c\sqrt{\frac{L^d n^{d/2-1}}{\epsilon^d}}.$$

Optimizing over the choice of ϵ, that is, setting

$$\epsilon = \left(\frac{cd\sqrt{L}}{4}\right)^{\frac{2}{d+2}} n^{\frac{d}{2(d+2)}}$$

gives

$$\mathbf{E}\|R_n\|_F \leq c_d \left(\frac{L}{n}\right)^{\frac{1}{d+2}}.$$

with

$$c_d = 2^{\frac{d-2}{d+2}} d^{\frac{2}{d+2}} + 2^{-\frac{2d}{d+2}} d^{-\frac{d}{d+2}}.$$