# Theoretical Statistics. Lecture 12.

**Peter Bartlett**

Uniform laws of large numbers: Bounding Rademacher complexity.

1. Metric entropy.
2. Canonical Rademacher and Gaussian processes

# Recall: Covering numbers

A **pseudometric** is like a metric, but we don't insist that $d(x, y) = 0$ implies $x = y$.

**Definition:** An $\epsilon$-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 $\epsilon$-covering number of $T$ is

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

A set $T$ is **totally bounded** if, for all $\epsilon > 0$, $N(\epsilon, T, d) < \infty$.
The function $\epsilon \mapsto \log N(\epsilon, T, d)$ is the **metric entropy** of $T$.
If $\lim_{\epsilon \to 0} \log N(\epsilon) / \log(1/\epsilon)$ exists, it is called the **metric dimension**.

## Covering numbers

Intuition: A $d$-dimensional set has metric dimension $d$. ($N(\epsilon) = \Theta(1/\epsilon^d)$.)

Example: $([0, 1]^d, l_\infty)$ has $N(\epsilon) = \Theta(1/\epsilon^d)$.

# Packing numbers

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

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

## Covering and packing numbers

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

Thus, the scaling of the covering and packing numbers is the same.

## Covering and packing numbers: Proof

For the first inequality, consider a minimal $\epsilon$-cover $\hat{T}$. Any two elements of a $2\epsilon$-packing of $T$ cannot be within $\epsilon$ of the same element of $\hat{T}$. (Otherwise, the triangle inequality shows that they are within $2\epsilon$ of each other.) Thus, there can be no more than one element of a $2\epsilon$ packing for each of the $N(\epsilon)$ elements of $\hat{T}$. That is, $M(2\epsilon) \leq N(\epsilon)$.

For the second inequality, consider an $\epsilon$-packing $\hat{T}$ of size $M(\epsilon)$. Since it is maximal, no other point $s \in T$ can be added for which some $t \in \hat{T}$ has $d(s,t) > \epsilon$. Thus, $\hat{T}$ is an $\epsilon$-cover. So the minimal $\epsilon$-cover has size $N(\epsilon) \leq M(\epsilon)$.

# Covering and packing numbers: Example

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

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

## Covering and packing numbers of a norm ball: Proof

Lower bound: Consider an $\epsilon$-cover $\{x_1, \ldots, x_N\}$ of size $N = N(\epsilon, B)$, and notice that

$$B \subseteq \bigcup_{i=1}^{N} (x_i + \epsilon B),$$

so $$\mathrm{vol}(B) \leq N(\epsilon, B)\mathrm{vol}(\epsilon B) = N(\epsilon, B)\epsilon^d \mathrm{vol}(B),$$

and hence $N(\epsilon, B) \geq 1/\epsilon^d$.

## Covering and packing numbers of a norm ball: Proof

Upper bound: Consider a maximal $\epsilon$-packing $\{x_1, \ldots, x_M\}$ of size $M = M(\epsilon, B)$. Since it's a packing, the balls $x_i + (\epsilon/2)B$ are disjoint. Each of these balls is contained in $(1 + \epsilon/2)B$. Thus,

$$\bigcup_{i=1}^{M} \left( x_i + \frac{\epsilon}{2} B \right) \subseteq (1 + \epsilon/2)B,$$

so

$$M \mathrm{vol}((\epsilon/2)B) \leq \mathrm{vol}((1 + \epsilon/2)B)$$

$$M \left( \frac{\epsilon}{2} \right)^d \mathrm{vol}(B) \leq \left( 1 + \frac{\epsilon}{2} \right)^d \mathrm{vol}(B).$$

and hence $N(\epsilon, B) \leq M(\epsilon, B) \leq (2/\epsilon + 1)^d$.

## Example: smoothly parameterized functions

Let $F$ be a parameterized class of functions,

$$F = \{f(\theta, \cdot) : \theta \in \Theta\}.$$

Let $\|\cdot\|_\Theta$ be a norm on $\Theta$ and let $\|\cdot\|_F$ be a norm on $F$. Suppose that the mapping $\theta \mapsto f(\theta, \cdot)$ is $L$-Lipschitz, that is,

$$\|f(\theta, \cdot) - f(\theta', \cdot)\|_F \leq L\|\theta - \theta'\|_\Theta.$$

Then $N(\epsilon, F, \|\cdot\|_F) \leq N(\epsilon/L, \Theta, \|\cdot\|_\Theta)$.

## Example: smoothly parameterized functions

A Lipschitz parameterization allows us to translates a cover of the parameter space into a cover of the function space.

Example: If $F$ is smoothly parameterized by a (compact set of) $d$ parameters, then $N(\epsilon, F) = O(1/\epsilon^d)$.

## Example: $1$-dimensional Lipschitz functions

Let $F$ be the set of $L$-Lipschitz functions mapping from $[0, 1]$ to $[0, 1]$. Then in the infinity norm $\|f\|_\infty = \sup_{x \in [0,1]} |f(x)|$,

$$\log N(\epsilon, F, \|\cdot\|_\infty) = \Theta(L/\epsilon).$$

Proof idea: form an $\epsilon$ grid of the y-axis, and an $\epsilon/L$ grid of the x-axis, and consider all functions that are piecewise linear on this grid, where all pieces have slopes $+L$ or $-L$. There are $1/\epsilon$ starting points, and for each starting point there are $2^{L/\epsilon}$ slope choices. It's easy to show that this set is an $O(\epsilon)$ packing and an $O(\epsilon)$ cover.

## Example: $d$-dimensional Lipschitz functions

Let $F_d$ be the set of $L$-Lipschitz functions (wrt $\|\cdot\|_\infty$) mapping from $[0,1]^d$ to $[0,1]$. Then

$$\log N(\epsilon, F_d, \|\cdot\|_\infty) = \Theta\left((L/\epsilon)^d\right).$$

Note the *exponential* dependence on the dimension.

# 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 $\epsilon_i$ are i.i.d. and uniform on $\{\pm 1\}$.

# 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.

# Canonical Rademacher and Gaussian Processes

Indeed:

$$G_\theta - G_{\theta'} = \langle g, \theta - \theta' \rangle,$$

which is $N(0, \|\theta - \theta'\|^2)$, and hence its moment generating function is equal to the upper bound.

$$R_\theta - R_{\theta'} = \langle \epsilon, \theta - \theta' \rangle,$$

which, by the bounded differences property, is sub-Gaussian with parameter $\|\theta - \theta'\|^2$.

# An aside: Orlicz norms

**Definition:** For $1 \le \alpha \le 2$, the $\alpha$-Orlicz norm of a random variable $X$ is

$$\|X\|_{\psi_\alpha} = \inf\left\{C > 0 : \mathbf{E}\exp\left(\frac{|X|^\alpha}{C^\alpha}\right) \le 2\right\}.$$

**Theorem:** There are constants $c_1, c_2$ such that, for all $X$ and all $t \ge 1$,

$$\Pr(|X| \ge t) \le 2\exp\left(-c_1 \frac{t^\alpha}{\|X\|_{\psi_\alpha}^\alpha}\right),$$

and conversely, $\Pr(|X| \ge t) \le c\exp(-t^\alpha / K^\alpha)$ implies $\|X\|_{\psi_\alpha} \le c_2 K$.

Sub-Gaussian means $\|X_\theta - X'_\theta\|_{\psi_2} \le L d(\theta, \theta')$.

# Canonical Gaussian and Rademacher processes

**Theorem:** For $T \subseteq \mathbb{R}^n$,

$$\mathbf{E} \sup_{\theta \in T} R_\theta \leq \sqrt{\frac{\pi}{2}} \mathbf{E} \sup_{\theta \in T} G_\theta \leq c\sqrt{\log n} \mathbf{E} \sup_{\theta \in T} R_\theta .$$

# Canonical Gaussian and Rademacher processes

Proof of first inequality:

$$\begin{aligned}
\mathbf{E}\sup_{\theta\in T} G_\theta &= \mathbf{E}\sup_{\theta\in T}\sum_{i=1}^n g_i\theta_i \\
&= \mathbf{E}\sup_{\theta\in T}\sum_{i=1}^n \epsilon_i|g_i|\theta_i \\
&\geq \mathbf{E}\sup_{\theta\in T}\sum_{i=1}^n \epsilon_i\mathbf{E}\left[|g_i|\right]\theta_i \\
&= \sqrt{\frac{2}{\pi}}\mathbf{E}\sup_{\theta\in T} R_\theta.
\end{aligned}$$

## Canonical Gaussian and Rademacher processes: Example

For $\Theta$ the $l_1$-ball in $\mathbb{R}^n$,

$$\mathbf{E} \sup_{\theta} \langle \epsilon, \theta \rangle = \mathbf{E} \|\epsilon\|_\infty = 1.$$

[where we've used the duality of $\ell_1$ and $\ell_\infty$ (equivalently, that Hölder's inequality is tight).] Also,

$$\mathbf{E} \sup_{\theta} \langle g, \theta \rangle = \mathbf{E} \|g\|_\infty \leq \sqrt{2 \ln n}.$$

The Gaussian and Rademacher complexities are a $\sqrt{\log n}$ factor apart in this case.

## Canonical Gaussian and Rademacher processes: Example

To see the last inequality, we generalize the Finite Lemma to the sub-Gaussian case:

**Lemma:** For $g$ with independent sub-Gaussian components,

$$\mathbf{E} \max_{a \in A} \langle g, a \rangle \leq \max_{a \in A} \|a\| \sqrt{2 \log |A|}.$$

In this case, $A = \{e_i : 1 \leq i \leq n\}$, so $\max_{a \in A} \|a\| = 1$ and $|A| = n$.

# Canonical Gaussian and Rademacher processes: Example

Proof:

$$\begin{aligned}
\exp\left(\lambda \mathbf{E} \max_{a \in A} \langle g, a \rangle\right) &\le \mathbf{E} \exp\left(\lambda \max_{a \in A} \langle g, a \rangle\right) \\
&= \mathbf{E} \max_{a \in A} \exp\left(\lambda \langle g, a \rangle\right) \\
&\le \sum_{a \in A} \mathbf{E} \exp\left(\lambda \langle g, a \rangle\right) \\
&\le |A| \exp\left(\lambda^2 R^2 / 2\right),
\end{aligned}$$

since $g_i$ is sub-gaussian (here, $R^2 = \max_{a \in A} \|a\|_2^2$). Picking $\lambda^2 = 2 \log |A| / R^2$ gives the result.