How we show uniform laws

- Show individual points converge
- Argue that set is not “too” large somehow

This lecture: understand how “large” sets are

Covering

Definition (Covering)

Let (T, ρ) be a metric space. A collection $\mathcal{N} = \{t_1, \ldots, t_N\}$ is an ϵ-*cover* if

$$\min_i \rho(t, t_i) \leq \epsilon \text{ for all } t \in T$$

Packing

Definition (Packing)

Let (T, ρ) be a metric space. A collection $\mathcal{M} = \{t_1, \ldots, t_M\}$ is a δ-*packing* if

$$\rho(t_i, t_j) > \delta \text{ for all } i \neq j.$$

Covering and packing numbers

Definition (Covering numbers)

The ϵ-*covering number* of a metric space (T, ρ) is

$$N(\epsilon; T, \rho) := \inf \{N \in \mathbb{N} \text{ s.t. } \exists \text{ an } \epsilon\text{-cover } t_1, \ldots, t_N\}$$

Definition (Packing numbers)

The δ-*packing number* of a metric space (T, ρ) is

$$M(\delta; T, \rho) := \sup \{M \in \mathbb{N} \text{ s.t. } \exists \text{ an} \delta\text{-packing } t_1, \ldots, t_M\}$$

Metric entropies

Definition (Entropies)

The *metric entropy* of a metric space (T, ρ) is $\log N(\epsilon; T, \rho)$. The *packing entropy* is $\log M(\epsilon; T, \rho)$

Proposition

For any metric space (T, ρ) *and* $\epsilon > 0$ *we have*

$$M(2\epsilon; T, \rho) \leq N(\epsilon; T, \rho) \leq M(\epsilon; T, \rho)$$

Example: Boolean hypercube

Let $T = \{0,1\}^d$ with metric $\rho(u,v) = \sum_{j=1}^d |u_j - v_j|$. Then there is a numerical constant $c > 0$ such that

$$c \cdot d \le \log N(d/4; T, \rho) \le d.$$

Example: norm ball, covering, and volume

Let $\|\cdot\|$ be any norm on $\mathbb{R}^d$ and $\mathbb{B} = \{x \in \mathbb{R}^d : \|x\| \le 1\}$ its unit ball. Then

$$\left(\frac{1}{\delta}\right)^d \le N(\delta; \mathbb{B}, \|\cdot\|) \le \left(1 + \frac{2}{\delta}\right)^d.$$

Example: Lipschitz functions on $[0, 1]$

Let $\mathcal{F} \subset \{f : [0, 1] \to \mathbb{R}\}$ be the 1-Lipschitz functions on $[0, 1]$ with $f(0) = 0$. Then

$$\log N(\delta; \mathcal{F}, \|\cdot\|_\infty) \asymp \frac{1}{\delta}$$

An application: concentration of i.i.d. sums of Lipschitz functions

Let $\ell : \Theta \times \mathcal{X} \to \mathbb{R}$ be 1-Lipschitz in θ, i.e.

$$|\ell(\theta, x) - \ell(\theta', x)| \le \left\|\theta - \theta'\right\|$$

and bounded with $\ell(\theta, x) \in [0, B]$.

Proposition

Let $\widehat{L}_n(\theta) = \frac{1}{n}\sum_{i=1}^n \ell(\theta; X_i)$. *Then*

$$\mathbb{P}\left(\sup_{\theta\in\Theta} |\widehat{L}_n(\theta) - L(\theta)| \ge t + \epsilon\right) \le N(\epsilon; \Theta, \|\cdot\|) \exp\left(-\frac{nt^2}{B^2}\right)$$

Concentration of i.i.d. sums of Lipschitz functions: picture

Concentration of i.i.d. sums of Lipschitz functions: proof

An application: matrix concentration

The matrix *operator norm* is

$$\|A\|_{\text{op}} = \sup_{x : \|x\|_2 \le 1} \|Ax\|_2$$

Suppose the matrix $A \in \mathbb{R}^{m \times n}$ has independent entries. What do we expect its operator norm to scale as?

Theorem

Let A_{ij} be independent σ^2-sub-Gaussian. There exists a numerical constant C such that

$$\mathbb{P}\left(\|A\|_{\text{op}} \ge C\sqrt{n} + C\sqrt{m} + Ct\right) \le 2e^{-t^2}.$$

Idea: Show that $u^T A v \approx 0$ with high probability, then cover.

Proof of concentration: discretization

Lemma

Let $\mathcal{N}_n, \mathcal{N}_m$ be ϵ-covers of the unit spheres in $\mathbb{R}^n$ and $\mathbb{R}^m$. Then

$$\max_{u \in \mathcal{N}_m, v \in \mathcal{N}_n} u^T A v \le \|A\|_{\rm op} \le \frac{1}{1 - 2\epsilon} \max_{u \in \mathcal{N}_m, v \in \mathcal{N}_n} u^T A v$$

Proof of concentration: sub-Gaussianity

Let $\mathcal{N}_n, \mathcal{N}_m$ be minimal $\frac{1}{4}$-covers of the unit spheres in $\mathbb{R}^n, \mathbb{R}^m$.

$$\mathbb{P}(\|A\|_{\text{op}} \geq \epsilon) \leq \mathbb{P}\left(\max_{u \in \mathcal{N}_m} \max_{v \in \mathcal{N}_n} u^T A v \geq \frac{\epsilon}{4}\right)$$

Proof of concentration: union bound

Sub-Gaussian processes and chaining

So far, we have seen

(i) Sub-Gaussian variables

(ii) Rademacher complexities

(iii) Covering numbers

Is there something that unifies them?

Sub-Gaussian process

Definition (Sub-Gaussian Process)

A collection of zero-mean random variables $\{X_\theta, \theta \in T\}$ is a *sub-Gaussian process* with respect to a metric ρ on T if

$$\mathbb{E}\left[e^{\lambda(X_\theta - X_{\theta'})}\right] \le \exp\left(\frac{\lambda^2 \rho(\theta, \theta')^2}{2}\right).$$

Example

Take $Z \sim \mathsf{N}(0, I_d)$ and $T = \mathbb{R}^d$, $\rho(\theta, \theta') = \|\theta - \theta'\|_2$, $X_\theta = \langle Z, \theta \rangle$

Sub-Gaussian process: symmetrized functions

Example

Let $\mathcal{F}$ be collection of $f : \mathcal{X} \to \mathbb{R}$, $\varepsilon_i \stackrel{\text{iid}}{\sim} \{\pm 1\}$, fix $x_1, \ldots, x_n$

$$Z_f := \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \varepsilon_i f(x_i)$$

Sub-Gaussian process: symmetrized functions

Example

Let $\ell : \Theta \times \mathcal{X} \to \mathbb{R}$ be B-Lipschitz, $\varepsilon_i \stackrel{\text{iid}}{\sim} \{\pm 1\}$, fix $x_1, \ldots, x_n$, set

$$Z_\theta := \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \varepsilon_i \ell(\theta, x_i)$$

Entropy integral

Question: Can we control Rademacher (or other complexities) by metric entropies?

Definition (Entropy integral)

Dudley's *entropy integral* is

$$J(D) := \int_0^D \sqrt{\log N(\epsilon; T, \rho)} d\epsilon.$$

Example

Lipschitz functions on $[0, 1]$ with $f(0) = 0$: $J(\infty) \lesssim \int_0^1 \epsilon^{-\frac{1}{2}} d\epsilon$

Entropy integral

Theorem (Dudley)

Let $\{X_\theta : \theta \in T\}$ *be a* ρ*-sub-Gaussian process with* $D \geq \sup_{\theta,\theta' \in T} \rho(\theta, \theta')$*. Then*

$$\mathbb{E}\left[\sup_{\theta,\theta' \in T} (X_\theta - X_{\theta'})\right] \lesssim \int_0^D \sqrt{\log N(\epsilon; T, \rho)} d\epsilon.$$

Example (Rademacher complexity of Lipschitz loss class)

Proof of entropy integral

Assume that process is *separable*, i.e. that exists set $T' \subset T$ with T' countable, $\sup_{\theta \in T'} X_\theta = \sup_{\theta \in T} X_\theta$

- Step 1. Construct a series of finer and finer discretizations

Proof of entropy integral

- Step 2. Construct projections (the chain)

Proof of entropy integral

- Step 3. Sum expected worst-case errors

Proof of entropy integral

- Step 4. Transform into integral

Example: VC Dimension

Let $\mathcal{F}$ be a class of Boolean functions with VC-dimension d. Then

$$\log N(\epsilon; \mathcal{F}, \|\cdot\|_{L^2(P_n)}) \lesssim d \log \frac{1}{\epsilon}$$

Proposition

We have $R_n(\mathcal{F}) \le C\sqrt{d/n}$ *and thus*

$$\mathbb{P}\left(\sup_{f \in \mathcal{F}} \left| \frac{1}{n} \sum_{i=1}^n f(X_i) - \mathbb{E}[f(X)] \right| \ge C\sqrt{\frac{d}{n}} + t\right) \le 2\exp(-nt^2).$$

Example: bounded Lipschitz functions

Let $\ell(\theta; x)$ be B-bounded and K-Lipschitz in θ, suppose $\log N(\epsilon; \Theta, \|\cdot\|) \le D \log \frac{1}{\epsilon}$. Let $\mathcal{F} = \{\ell(\theta; \cdot) \mid \theta \in \Theta\}$. Then

$$R_n(\mathcal{F}) \lesssim \frac{BKD}{\sqrt{n}}$$

Multiclass classification

Consider k-class classification problem,

$$\theta = \begin{bmatrix} \theta^1 & \theta^2 & \cdots & \theta^k \end{bmatrix} \in \mathbb{R}^{d \times k}$$

Let margin $s = \theta^T x \in \mathbb{R}^k$, loss $\phi : \mathbb{R}^k \to \mathbb{R}$ of form

$$\ell(\theta; x, y) = \phi(\Pi_y s) = \phi(\Pi_y \theta^T x)$$

for some "labeling" matrix Π_y

Rademacher complexity and generalization for multiclass

Rademacher complexity and generalization for multiclass