Group Study and Seminar Series (Summer 20), Week 5

Uniform Law of Large Number

Presenters: Qingyan Meng

The Chinese University of Hong Kong, Shenzhen, China

July 16, 2020

Mainly based on:

Wainwright, M. J. (2019). High-dimensional statistics: A non-asymptotic viewpoint (Vol. 48). Cambridge University Press.

Shalev-Shwartz, S., & Ben-David, S. (2014). Understanding machine learning: From theory to algorithms. Cambridge university press.

Outline

Some Concepts of the Statistical Learning Framework

- Training data S
  - $S = \{(x_1, y_1), \ldots, (x_n, y_n)\}$ contains n i.i.d. copies of a random variable (X,Y) with distribution $\mathcal{D}$, where $y \in \mathcal{Y} \subseteq \mathbb{R}$.
  - Define $\mathcal{Z} = (\mathcal{X}, \mathcal{Y})$ for later use.
- Hypothesis class $\mathcal{H}$

  $\mathcal{H} = \{h : \mathcal{X} \to \mathcal{Y}\}$ is a class of predictors.
- Loss function ℓ

  $\ell : \mathcal{H} \times \mathcal{Z} \to \mathbb{R}^+$ measures the error of $h \in \mathcal{H}$ with respect to a sample $z \in \mathcal{Z}$.
- An example: linear regression

  $\mathcal{X} = \mathbb{R}^d$, $\mathcal{Y} = \mathbb{R}$, $\mathcal{H} = \left\{x \mapsto \langle w, x\rangle : w \in \mathbb{R}^d\right\}$, $\ell(h, (x, y)) = (h(x) - y)^2$.

Empirical Risk Minimization

- Define $\mathcal{F} = \ell \circ \mathcal{H} = \{z \mapsto \ell(h, z) : h \in \mathcal{H}, z \in \mathcal{Z}\}$.
- True error (expected error)
  $L_{\mathcal{D}}(h) \triangleq L_{\mathcal{D}}(f) \triangleq \underset{z \sim \mathcal{D}}{\mathbb{E}}[f(z)]$, where $f \in \mathcal{F}$ and $f(\cdot) = \ell(h, \cdot)$.
- Empirical error
  $L_S(h) \triangleq L_S(f) \triangleq \frac{1}{m} \sum_{i=1}^{m} f(z_i)$, where m is the number of samples in S.
- Fix the hypothesis class $\mathcal{H}$, given the training data S, the empirical risk minimization (ERM) method is defined as

$$\widehat{h} \in \underset{h \in \mathcal{H}}{\operatorname{argmin}} \frac{1}{m} \sum_{i=1}^{m} \ell(h, z_i).$$

Error Analysis

- Suppose we get a predictor $\widehat{h}$ by ERM method. How good this predictor is?
  - Or, how big the gap between $L_D(\widehat{h})$ and $L_D(h^*)$ is? Here $h^* \in \operatorname*{argmin}_{h \in \mathcal{H}} L_D(h)$.
  - It is our final goal today.
- $L_D(\widehat{h}) - L_D(h^*) = \underbrace{L_D(\widehat{h}) - L_S(\widehat{h})}_{(1)} + \underbrace{L_S(\widehat{h}) - L_S(h^*)}_{(2)} + \underbrace{L_S(h^*) - L_D(h^*)}_{(3)}$.
  - (2): non-positive;
  - (3): zero-mean, and its fluctuations can be controlled with the tail bounds;
  - (1): not zero-mean because $\widehat{h}$ depends on samples, challenging to control.
- We need to bound (1).

Problem Introduction

- To bound (1), define $Rep(\mathcal{F}, S) = \sup_{h \in \mathcal{H}} (L_D(h) - L_S(h))$, then $L_D(\widehat{h}) - L_S(\widehat{h}) \leq Rep(\mathcal{F}, S)$, so:

$$\mathop{\mathbb{E}}_{S}(L_D(\widehat{h}) - L_D(h^*)) \leq \mathop{\mathbb{E}}_{S}[Rep(\mathcal{F}, S)],$$

and

$$L_D(\widehat{h}) - L_D(h^*) \leq Rep(\mathcal{F}, S) + \epsilon(\delta, m)$$

with probability of at least $1 - \delta$ over the choice of S, for each $\delta \in (0, 1)$, where ϵ is gotten by tail bounds.

- Can we bound $\mathop{\mathbb{E}}_{S}[Rep(\mathcal{F}, S)]$ or bound $Rep(\mathcal{F}, S)$ with high probability?

Outline

Problem Setting

- Consider the task:
  - $\mathcal{X} = \mathbb{R}$, and $\mathcal{Y} = \emptyset$.
  - $\ell(\theta, x) = \mathbb{I}(x \leq \theta)$.
  - $\mathcal{F} = \{x \to \mathbb{I}(x \leq \theta), \theta \in \mathbb{R}\}$.
- Try to upper bound $\underset{S}{\mathbb{E}}[Rep(\mathcal{F}, S)]$.
  - Method: introduce "ghost" variables $\{X_1^{'}, X_2^{'}, \cdots, X_m^{'}\}$, which are independent copies of X, to replace X.
  - Procedure: the next page.

Procedure for Bounding $\mathop{\mathbb{E}}_{S}[Rep(\mathcal{F}, S)]$

$$
\begin{aligned}
&\mathop{\mathbb{E}}_{S}[Rep(\mathcal{F}, S)] \\
&\triangleq \mathop{\mathbb{E}}_{\{X_i\}}\left[\sup_{\theta}\left(\mathbb{E}\mathbb{I}(X \le \theta) - \frac{1}{m}\sum_{i=1}^{m}\mathbb{I}(X_i \le \theta)\right)\right] \\
&= \mathop{\mathbb{E}}_{\{X_i\}}\left[\sup_{\theta}\left(\mathop{\mathbb{E}}_{\{X_i^{'}\}}\left[\frac{1}{m}\sum_{i=1}^{m}\mathbb{I}(X_i^{'} \le \theta)\right] - \frac{1}{m}\sum_{i=1}^{m}\mathbb{I}(X_i \le \theta)\right)\right] \\
&\le \mathop{\mathbb{E}}_{\{X_i, X_i^{'}\}}\left[\sup_{\theta}\left(\frac{1}{m}\sum_{i=1}^{m}\left[\mathbb{I}(X_i^{'} \le \theta) - \mathbb{I}(X_i \le \theta)\right]\right)\right] \\
&= \mathop{\mathbb{E}}_{\{X_i, X_i^{'}, \sigma_i \sim Rad()\}}\left[\sup_{\theta}\left(\frac{1}{m}\sum_{i=1}^{m}\sigma_i\left[\mathbb{I}(X_i^{'} \le \theta) - \mathbb{I}(X_i \le \theta)\right]\right)\right] \\
&= (next\ page)
\end{aligned}
$$

Procedure for Bounding $\mathop{\mathbb{E}}_{S}[Rep(\mathcal{F}, S)]$

$$
\begin{aligned}
&(last\ page)\\
&\leq \mathop{\mathbb{E}}_{\{X_i', \sigma_i\}} \left[\sup_{\theta} \left(\frac{1}{m} \sum_{i=1}^{m} \sigma_i \left[\mathbb{I}(X_i' \leq \theta) \right] \right) \right] + \mathop{\mathbb{E}}_{\{X_i, \sigma_i\}} \left[\sup_{\theta} \left(\frac{1}{m} \sum_{i=1}^{m} \sigma_i \left[\mathbb{I}(X_i \leq \theta) \right] \right) \right] \\
&= 2 \cdot \mathop{\mathbb{E}}_{\{X_i, \sigma_i\}} \left[\sup_{\theta} \left(\frac{1}{m} \sum_{i=1}^{m} \sigma_i \left[\mathbb{I}(X_i \leq \theta) \right] \right) \right] \\
&= 2 \cdot \mathop{\mathbb{E}}_{\{X_i\}} \left[\mathop{\mathbb{E}}_{\{\sigma_i\}} \left[\sup_{\theta} \left(\frac{1}{m} \sum_{i=1}^{m} \sigma_i \left[\mathbb{I}(X_i \leq \theta) \right] \right) \middle| \{X_i\} \right] \right] \\
&\triangleq 2 \cdot \mathop{\mathbb{E}}_{\{X_i\}} \left[R(\mathcal{F} \circ S) \right]
\end{aligned}
$$

Procedure for Bounding $\underset{S}{\mathbb{E}}[Rep(\mathcal{F}, S)]$

- Define $R(\mathcal{F} \circ S) = \frac{1}{m} \underset{\{\sigma_i\}}{\mathbb{E}} \left[\sup_{f \in \mathcal{F}} \left(\sum_{i=1}^{m} \sigma_i f(z_i) \right) \right]$ as Rademacher complexity of $\mathcal{F}$ with respect to S, where $z_i = (x_i, y_i)$ is a sample in S.
- Go back to the problem:

$$
\begin{aligned}
& 2 \underset{\{X_i\}}{\mathbb{E}} \left[R(\mathcal{F} \circ S) \right] \\
& = 2 \underset{\{X_i\}}{\mathbb{E}} \left[\underset{\{\sigma_i\}}{\mathbb{E}} \left[\sup_{\theta \in \mathbb{R}} \left(\frac{1}{m} \sum_{i=1}^{m} \sigma_i \left[\mathbb{I}(X_i \leq \theta) \right] \right) \middle| \{X_i\} \right] \right] \\
& = 2 \underset{\{X_i\}}{\mathbb{E}} \left[\underset{\{\sigma_i\}}{\mathbb{E}} \left[\max_{\theta \in \{\theta_1, \cdots, \theta_{m+1}\}} \left(\frac{1}{m} \sum_{i=1}^{m} \sigma_i \left[\mathbb{I}(X_i \leq \theta) \right] \right) \middle| \{X_i\} \right] \right] \\
& \leq 2 \frac{\sqrt{2 \log(m+1)}}{\sqrt{m}} \quad \leftarrow \textit{a property of Rademacher complexity, introduced later}
\end{aligned}
$$

Outline

Outline

Definition

- Define Rademacher complexity of $\mathcal{F}$ with respect to S:

$$R(\mathcal{F} \circ S) \triangleq \frac{1}{m} \mathop{\mathbb{E}}_{\{\sigma_i\}} \left[\sup_{f \in \mathcal{F}} \left(\sum_{i=1}^{m} \sigma_i f(z_i) \right) \right],$$

  where $z_i = (x_i, y_i)$ is a sample in S, m is the number of samples, σ_i is a random variable such that $\sigma_i = 1$ w.p. $\frac{1}{2}$ and $\sigma_i = -1$ w.p. $\frac{1}{2}$.

- More generally, given a set of vectors $A \subseteq \mathbb{R}^m$, define Rademacher complexity of A:

$$R(A) \triangleq \frac{1}{m} \mathop{\mathbb{E}}_{\{\sigma_i\}} \left[\sup_{a \in A} \left(\sum_{i=1}^{m} \sigma_i a_i \right) \right] = \frac{1}{m} \mathop{\mathbb{E}}_{\sigma} \left[\sup_{a \in A} (< \sigma, a >) \right].$$

Theorem

- **Lemma 1.** $\underset{S}{\mathbb{E}}\left[Rep(\mathcal{F}, S)\right] \leq 2\underset{S}{\mathbb{E}}\left[R(\mathcal{F} \circ S)\right]$.
  - The proof procedure is almost the same as the one in the section "Introducing Rademacher Complexity With a Simple Example", but a little more complicated.

Theorem

- **Theorem 2.** For any $h \in \mathcal{H}$,

$$\mathop{\mathbb{E}}_{S}\left[L_D(\widehat{h}) - L_D(h)\right] \leq 2\mathop{\mathbb{E}}_{S}\left[R(\mathcal{F} \circ S)\right].$$

Furthermore, for $\forall \delta \in (0,1)$, with probability of at least $1-\delta$ over the choice of S,

$$L_D(\widehat{h}) - L_D(h^*) \leq 2\mathop{\mathbb{E}}_{S}\left[R(\mathcal{F} \circ S)\right]/\delta.$$

  - Remind that $\widehat{h} \in \operatorname*{argmin}_{h \in \mathcal{H}} L_S(h)$ and $h^* \in \operatorname*{argmin}_{h \in \mathcal{H}} L_D(h)$.
  - The first inequality follows because $L_D(h) = \mathop{\mathbb{E}}_{S} L_S(h) \geq \mathop{\mathbb{E}}_{S} L_S(\widehat{h})$ for $\forall h$ and **Lemma 1**.
  - The second inequality follows from the first inequality by relying on Markov's inequality.

Theorem

- **Theorem 3.** Assume that for all z and $h \in \mathcal{H}$ we have that $|\ell(h, z)| \leq c$. For $\forall \delta \in (0, 1)$, with probability of at least $1 - \delta$ over the choice of S,

$$L_{\mathcal{D}}(h) - L_S(h) \leq 2 \mathop{\mathbb{E}}_{S'} R\left(\mathcal{F} \circ S'\right) + c\sqrt{\frac{2\ln(2/\delta)}{m}},$$

$$L_{\mathcal{D}}(h) - L_S(h) \leq 2R\left(\mathcal{F} \circ S\right) + 4c\sqrt{\frac{2\ln(4/\delta)}{m}},$$

$$L_{\mathcal{D}}(\widehat{h}) - L_D(h) \leq 2R\left(\mathcal{F} \circ S\right) + 5c\sqrt{\frac{2\ln(8/\delta)}{m}}.$$

  - Remind that $\widehat{h} \in \mathop{\operatorname{argmin}}_{h \in \mathcal{H}} L_S(h)$.
  - We will prove the first inequality. Please refer to Theorem 26.5 in [2] for the others.

Proof of Theorem 3

- Recall the **bounded differences inequality** we have learned in week 2 from Richard: Let $f : \mathbb{R}^m \to \mathbb{R}$ be a function of m variables such that $\left|f(x) - f\left(x^{\setminus k}\right)\right| \leq b$ for some $b > 0$ for all $x, x' \in \mathbb{R}^m$, then with probability of at least $1 - \delta$ we have

$$|f(X) - \mathbb{E}[f(X)]| \leq b\sqrt{\ln\left(\frac{2}{\delta}\right) m/2}.$$

- For the first inequality, note that $\mathrm{Rep}(\mathcal{F}, S) = \sup_{h \in \mathcal{H}} (L_{\mathcal{D}}(h) - L_S(h))$ satisfies the bounded differences condition with the constant $2c/m$, then

$$\mathrm{Rep}(\mathcal{F}, S) \leq \mathbb{E}\,\mathrm{Rep}(\mathcal{F}, S) + c\sqrt{\frac{2\ln(2/\delta)}{m}} \leq 2\mathop{\mathbb{E}}_{S'} R(\ell \circ \mathcal{H} \circ S') + c\sqrt{\frac{2\ln(2/\delta)}{m}}.$$

Remark

- **Why we introduce Rademacher complexity?**
  We need to derive generalization bound. From **Lemma 1** to **Theorem 3**, we see that those generalization errors can be bounded by something related to Rademacher complexity, so we can bound Rademacher complexity instead.
- **Is Rademacher complexity easier to bound?**
  Yes. There are many properties of Rademacher complexity to use, which will be introduced in the next subsection.

Outline

Properties

- **Property 4.** For any $A \subset \mathbb{R}^m$, scalar $c \in \mathbb{R}$, and vector $\mathbf{a}_0 \in \mathbb{R}^m$, we have

$$R\left(\{c\mathbf{a} + \mathbf{a}_0 : \mathbf{a} \in A\}\right) \leq |c| R(A).$$

  That is, linear transformation linearly changes the Rademacher complexity of a set.

- **Property 5.** Let A be a subset of $\mathbb{R}^m$ and let $A' = \left\{\sum_{j=1}^{N} \alpha_j \mathbf{a}^{(j)} : N \in \mathbb{N}, \forall j, \mathbf{a}^{(j)} \in A,\right.$

$$\left.\alpha_j \geq 0, \|\boldsymbol{\alpha}\|_1 = 1\right\}. \text{ Then }, R\left(A'\right) = R(A).$$

  That is, the convex hull of A has the same Rademacher complexity as A.

Properties

- **Property 6. (Massart)** Let $A = \{\mathbf{a}_1, \ldots, \mathbf{a}_N\}$ be a finite set of vectors in $\mathbb{R}^m$. Define $\overline{\mathbf{a}} = \frac{1}{N}\sum_{i=1}^{N} \mathbf{a}_i$. Then,

$$R(A) \leq \max_{\mathbf{a}\in A} \|\mathbf{a} - \overline{\mathbf{a}}\| \frac{\sqrt{2\log(N)}}{m},$$

  or,

$$R(A) \leq \max_{\mathbf{a}\in A} \|\mathbf{a}\| \frac{\sqrt{2\log(N)}}{m}.$$

- **Property 7. (Contraction Inequality)** For each $i \in [m]$, let $\phi_i : \mathbb{R} \to \mathbb{R}$ be a $\rho-$Lipschitz function, let $\phi(a) \triangleq (\phi_1(a_1), \ldots, \phi_m(a_m))$ and $\phi \circ A \triangleq \{\phi(a) : a \in A\}$, then

$$R(\phi \circ A) \leq \rho R(A).$$

Proof for Property 6 (The Second Inequality)

- Let $W_a = \frac{1}{m}\sum_{i=1}^{m} \sigma_i a_i$, then $\mathbb{R}(\mathcal{A}) = \mathbb{E}\left[\sup_{a\in\mathcal{A}} W_a\right]$.
- $\exp\left(t\mathbb{E}\left[\sup_{a\in\mathcal{A}} W_a\right]\right) \leq \mathbb{E}\left[\exp\left(t \sup_{a\in\mathcal{A}} W_a\right)\right] = \mathbb{E}\left[\sup_{a\in\mathcal{A}} \exp\left(tW_a\right)\right] \leq \sum_{a\in\mathcal{A}} \mathbb{E}\left[\exp\left(tW_a\right)\right].$
- σ_i is a bounded random variable, thus is sub-Gaussian with parameter $2^2/4 = 1$. So, W_a is sub-Gaussian with parameter $\max_{\mathbf{a}\in A} \|\mathbf{a}\|^2/m^2$. (P24 in Richard's slides)
- By the definition of sub-Gaussian,

$$\mathbb{E}\left[\exp\left(tW_a\right)\right] \leq \exp\left(\frac{t^2 \max_{\mathbf{a}\in A} \|\mathbf{a}\|^2}{2m^2}\right)$$

.

- Plugging the above formula into the overall bound, taking logs, and optimizing over t yields the result.

Properties

- **Property 8. (example)** Let $\mathcal{H}_2 = \{\mathrm{x} \mapsto \langle \mathbf{w}, \mathrm{x} \rangle : \|\mathbf{w}\|_2 \leq 1\}$ and $S = (\mathrm{x}_1, \ldots, \mathrm{x}_m)$ be vectors in a Hilbert space, then

$$R\left(\mathcal{H}_2 \circ S\right) \leq \frac{\max_i \|\mathbf{x}_i\|_2}{\sqrt{m}}.$$

- **Property 9. (example)** Let $\mathcal{H}_1 = \{\mathrm{x} \mapsto \langle \mathbf{w}, \mathrm{x} \rangle : \|\mathbf{w}\|_1 \leq 1\}$ and $S = (\mathrm{x}_1, \ldots, \mathrm{x}_m)$ be vectors in $\mathbb{R}^n$, then

$$R\left(\mathcal{H}_1 \circ S\right) \leq \max_i \|\mathbf{x}_i\|_\infty \sqrt{\frac{2\log(2n)}{m}}.$$

Proof for Property 8

$$
\begin{aligned}
mR(\mathcal{H}_2 \circ S) &= \mathbb{E}_{\boldsymbol{\sigma}}\left[\sup_{\mathbf{w}:\|\mathbf{w}\|\leq 1}\sum_{i=1}^{m}\sigma_i\left\langle \mathbf{w}, \mathbf{x}_i\right\rangle\right] = \mathbb{E}_{\boldsymbol{\sigma}}\left[\sup_{\mathbf{w}:\|\mathbf{w}\|\leq 1}\left\langle \mathbf{w}, \sum_{i=1}^{m}\sigma_i\mathbf{x}_i\right\rangle\right] \\
&\leq \mathbb{E}_{\boldsymbol{\sigma}}\left[\left\|\sum_{i=1}^{m}\sigma_i\mathbf{x}_i\right\|_2\right] = \mathbb{E}_{\boldsymbol{\sigma}}\left[\left(\left\|\sum_{i=1}^{m}\sigma_i\mathbf{x}_i\right\|_2^2\right)^{\frac{1}{2}}\right] \\
&\leq \left(\mathbb{E}_{\boldsymbol{\sigma}}\left[\left\|\sum_{i=1}^{m}\sigma_i\mathbf{x}_i\right\|_2^2\right]\right)^{\frac{1}{2}} = \left(\sum_{i\neq j}\left\langle \mathbf{x}_i, \mathbf{x}_j\right\rangle \mathbb{E}_{\boldsymbol{\sigma}}\left[\sigma_i\sigma_j\right] + \sum_{i=1}^{m}\left\langle \mathbf{x}_i, \mathbf{x}_i\right\rangle \mathbb{E}_{\boldsymbol{\sigma}}\left[\sigma_i^2\right]\right)^{\frac{1}{2}} \\
&= \left(\sum_{i=1}^{m}\|\mathbf{x}_i\|_2^2\right)^{\frac{1}{2}} \leq \left(m\max_i\|\mathbf{x}_i\|_2^2\right)^{\frac{1}{2}}.
\end{aligned}
$$

Proof for Property 9

(The first inequality follows from Holder's and Jensen's inequality.)

$$
\begin{aligned}
mR\left(\mathcal{H}_1 \circ S\right) &\leq \mathop{\mathbb{E}}_{\boldsymbol{\sigma}}\left[\left\|\sum_{i=1}^{m}\sigma_i \mathbf{x}_i\right\|_{\infty}\right] = \mathop{\mathbb{E}}_{\boldsymbol{\sigma}}\left[\max_{j\in[n]}\left|\sum_{i=1}^{m}\sigma_i \mathbf{x}_{ij}\right|\right] \\
&\triangleq \mathop{\mathbb{E}}_{\boldsymbol{\sigma}}\left[\max_{j\in[n]}|<\boldsymbol{\sigma}, \boldsymbol{v}_j>|\right] \quad \leftarrow \ let\ \boldsymbol{v}_j = (\mathbf{x}_{1j}, \cdots, \mathbf{x}_{mj})^T \\
&= \mathop{\mathbb{E}}_{\boldsymbol{\sigma}}\left[\max_{j\in[n]}[\max(<\boldsymbol{\sigma}, \boldsymbol{v}_j>, <\boldsymbol{\sigma}, -\boldsymbol{v}_j>)]\right] \\
&\triangleq \mathop{\mathbb{E}}_{\boldsymbol{\sigma}}\left[\max_{\boldsymbol{v}\in V}<\boldsymbol{\sigma}, \boldsymbol{v}>\right] \quad \leftarrow \ let\ V = (\mathbf{v}_1, \cdots, \mathbf{v}_n, -\mathbf{v}_1, \cdots, -\mathbf{v}_n)^T \\
&= mR(V).
\end{aligned}
$$

Note that $\max_{\boldsymbol{v}\in V}||\mathbf{v}-\hat{\mathbf{v}}||_2 = \max_{\boldsymbol{v}\in V}||\mathbf{v}||_2 \leq \sqrt{m}\ \max_i||\mathbf{x}_i||_{\infty}$, we get the result by **Property 6**.

Outline

Example 1

- See the example in the section "Introducing Rademacher Complexity With a Simple Example".
- In this example, we transfer $R(\mathcal{F} \circ S) = R(\{x \to \mathbb{I}(x \leq \theta), \theta \in \mathbb{R}\} \circ S)$ to Rademacher complexity of a finite set, then use **Property 6** to get the result.

Example 2: Problem Setting

Consider the hard-SVM algorithm:

- Consider a distribution $\mathcal{D}$ over $\mathcal{X} \times \{\pm 1\}$, such that $\exists \mathbf{w}^*$ with $\mathbb{P}_{(\mathbf{x},y)\sim\mathcal{D}}\left[y \left\langle \mathbf{w}^\star, \mathbf{x} \right\rangle \geq 1\right] = 1$ (separability assumption holds), and $||\mathbf{x}||_2 \leq R$ w.p. 1, where $\mathbf{x} \in \mathcal{X}$.
- Consider the hard-SVM problem:

$$\underset{\mathbf{w}}{\operatorname{argmin}} \; ||\mathbf{x}||_2^2 \quad s.t. \; y_i \left\langle \mathbf{w}, x_i \right\rangle \geq 1, \forall i.$$

- Define w_s is the output of this problem. Please give an upper bound on

$$\mathbb{P}_{(\mathbf{x},y)\sim\mathcal{D}}\left[y \neq \operatorname{sign}\left(\left\langle \mathbf{w}_S, \mathbf{x} \right\rangle\right)\right].$$

Example 2: Deriving bound by Rademacher Complexity

- Let $B = ||\mathbf{w}^*||_2$ and consider the hypothesis class $\mathcal{H} = \{\mathbf{w} : ||\mathbf{w}||_2 \leq B\}$. By the algorithm of hard-SVM, we know that $\mathbf{w}_s \in \mathcal{H}$.
- Consider the loss function to be ramp loss

$$\ell(\mathbf{w}, (\mathbf{x}, y)) = \min\{1, \max\{0, 1 - y\langle \mathbf{w}, \mathbf{x}\rangle\}\},$$

  then $|\ell| \leq 1$ and ℓ is 1-Lipschitz.
- By **Property 8**, $R(\mathcal{H} \circ S) \leq \frac{BR}{\sqrt{m}}$;
  and then by **Property 7**, $R(\ell \circ \mathcal{H} \circ S) \leq \frac{BR}{\sqrt{m}}$.

Example 2: Deriving bound by Rademacher Complexity

- Then by **Theorem 3**,

$$L_D(\mathbf{w}_s) - L_S(\mathbf{w}_s) \leq 2\mathop{\mathbb{E}}_{S'} R\left(\ell \circ \mathcal{H} \circ S'\right) + 1 \cdot \sqrt{\frac{2\ln(2/\delta)}{m}}$$

  with probability of at least $1-\delta$ over the choice of S.
- By the algorithm of hard-SVM, $L_S(\mathbf{w}_s) = 0$.
- By the definition of ramp loss, we can see

$$\mathop{\mathbb{P}}_{(\mathbf{x},y)\sim\mathcal{D}}\left[y \neq \operatorname{sign}\left(\langle \mathbf{w}_S, \mathbf{x}\rangle\right)\right] \leq L_D(\mathbf{w}_s).$$

- Above all,

$$\mathop{\mathbb{P}}_{(\mathbf{x},y)\sim\mathcal{D}}\left[y \neq \operatorname{sign}\left(\langle \mathbf{w}_S, \mathbf{x}\rangle\right)\right] \leq 2\frac{BR}{\sqrt{m}} + \sqrt{\frac{2\ln(2/\delta)}{m}}.$$

Example 3: Problem Setting

Consider a feed-forward neural network:

- A feed-forward neural network with depth ι ($\iota - 1$ hidden layers) is given by the function $f_{nn}^{\iota} : \mathbb{R}^d \to \mathbb{R}$ defined as

$$f_{nn}^{(\iota)}(x) := l^{(\iota)} \circ \cdots \circ l^{(1)}(x) \equiv l^{(\iota)} \left(\cdots l^{(2)} \left(l^{(1)}(x) \right) \cdots \right),$$

  where each layer $l^{(k)} : \mathbb{R}^{d_{k-1}} \to \mathbb{R}^{d_k}$ is a map:

$$l^{(k)}(x) := \sigma^{(k)} \left(\mathbf{w}^{(k)} x + b^{(k)} \right), \ \sigma^{(k)} \ is \ \lambda - Lipschitz.$$

- Calculate the Rademacher complexity for the network class:

$$\mathcal{A}_{nn}^{(\iota)} := \left\{ x \in \mathbb{R}^d \to f_{nn}^{(\iota)}(x) : \left\| \mathbf{w}^{(k)} \right\|_{\infty} \leq \omega, \left\| b^{(k)} \right\|_{\infty} \leq \beta \ \forall k, \sigma^{(\iota)}(x) = x \right\}.$$

Example 3: Solution

- Let $\sigma : \mathbb{R} \to \mathbb{R}$ be γ-Lipschitz. Define

$$\mathcal{L}' := \left\{ x \in \mathbb{R}^d \to \sigma \left(\sum_{j=1}^{m} w_j l_j(x) + b \right) \in \mathbb{R} : |b| \leq \beta, \|w\|_1 \leq \omega, l_1, \ldots, l_m \in \mathcal{L} \right\},$$

where $\mathcal{L}$ is a class of functions from $\mathbb{R}^d$ to $\mathbb{R}$ that includes the zero function. Let us first prove

$$\mathrm{R}\left(\mathcal{L}' \circ \{x_1, \ldots, x_n\}\right) \leq \gamma \left(\frac{\beta}{\sqrt{n}} + 2\omega \, \mathrm{R}\left(\mathcal{L} \circ \{x_1, \ldots, x_n\}\right) \right).$$

- To prove this, define

$$\mathcal{F} := \left\{ x \in \mathbb{R}^d \to \textstyle\sum_{i=1}^{m} w_j l_j(x) \in \mathbb{R} : \|w\|_1 \leq \omega, l_1, \ldots, l_m \in \mathcal{L} \right\}$$
$$\mathcal{G} := \left\{ x \in \mathbb{R}^d \to b \in \mathbb{R} : |b| \leq \beta \right\}$$

.

Example 3: Solution

Proof Sketch

- Firstly, get $\mathrm{R}\left(\mathcal{L}' \circ \{x_1, \ldots, x_n\}\right) \leq \gamma \left(\mathrm{Rad}\left(\mathcal{F} \circ \{x_1, \ldots, x_n\}\right) + \mathrm{Rad}\left(\mathcal{G} \circ \{x_1, \ldots, x_n\}\right)\right)$ by **Property 6**.
- Secondly, get $\mathrm{R}\left(\mathcal{F} \circ \{x_1, \ldots, x_n\}\right) \leq \omega\, \mathrm{R}\left(\mathrm{conv}(\mathcal{L} - \mathcal{L}) \circ \{x_1, \ldots, x_n\}\right)$, where $\mathcal{L} - \mathcal{L} \triangleq \{l - l' : l \in \mathcal{L}, l' \in \mathcal{L}\}$, by **Property 4** and the condition that 0 zero function is in $\mathcal{L}$.
- Then, get $\omega\, \mathrm{R}\left(\mathrm{conv}(\mathcal{L} - \mathcal{L}) \circ \{x_1, \ldots, x_n\}\right) = 2\omega\, \mathrm{R}\left(\mathcal{L} \circ \{x_1, \ldots, x_n\}\right)$ by **Property 5** and symmetry.
- Finally, get $n\, \mathrm{R}\left(\mathcal{G} \circ \{x_1, \ldots, x_n\}\right)) = \mathbf{E} \sup_{b:|b| \leq \beta} b \sum_{i=1}^n \sigma_i \leq \beta \mathbf{E} \left|\sum_{i=1}^n \sigma_i\right| \leq \beta \sqrt{n}$ by Jensen's inequality.

Example 3: Solution

With $\mathrm{R}\left(\mathcal{L}' \circ \{x_1, \ldots, x_n\}\right)$, let's derive $\mathrm{R}\left(\mathcal{A}_{nn}^{(\iota)} \circ \{x_1, \ldots, x_n\}\right)$.

- Use the result for $\mathrm{R}\left(\mathcal{L}' \circ \{x_1, \ldots, x_n\}\right)$ recursively for each layer (note that $\gamma = 1$ for the last layer and $\gamma = \lambda$ for the others) and **Property 7**, we find

$$\mathrm{R}\left(\mathcal{A}_{nn}^{(\iota)} \circ \{x_1, \ldots, x_n\}\right) \leq \frac{\beta}{\sqrt{n}} + 2\omega \left(\frac{\beta\lambda}{\sqrt{n}} \sum_{k=0}^{\iota-3} (2\omega\lambda)^k + (2\omega\lambda)^{\iota-2} \mathrm{R}\left(\mathcal{H}_1 \circ \{x_1, \ldots, x_n\}\right) \right),$$

  where $\mathcal{H}_1 = \{\mathrm{x} \mapsto \langle \mathbf{w}, \mathrm{x} \rangle : \|\mathbf{w}\|_1 \leq 1\}$.

- Then use **Property 9** to get the result.

Example 3: Comments

- Please refer to lecture notes 3 of reference [4] for detailed proof.
- This result cannot explain the practice and the success of deep learning. For example, in the limit of an infinite number of layers, it would have to be $2\omega\lambda < 1$, which is a restrictive requirement not needed in practice).
- Please refer to other papers for more reasonable bound under other assumptions. Also see Appendix B of reference [3] for the usage of Rademacher complexity in deriving network's generalization bound.

Outline

Covering Number: Motivation

- In **Property 9** and **Example 1**, instead of calculating Rademacher complexity directly, we solve equivalent problems where only finite hypothesis classes are involved and **Property 6** can be used to easily derive bound.
- We can apply the same idea in a general sense. For a set with infinity many points, we can isolate finitely points of interest, bound the Rademacher complexity of this finite subset, and bound the difference between Rademacher complexity of the original set and the new finite set.
- Covering number is the concept to measure cardinality of the finite set of interest.

Covering Number: Definition

- **Definition. Covering Number**
  Suppose $\mathcal{A} \subseteq \mathbb{R}^m$ and is equiped with a metric ρ. The set $\mathcal{C} \subseteq \mathcal{A}$ is a ε -cover of $(\mathcal{A}, \rho)$ if for every $x \in \mathcal{A}$ there exists $y \in \mathcal{C}$ such that $\rho(x, y) \leq \varepsilon$. The set $\mathcal{C} \subseteq \mathcal{A}$ is a minimal ε -cover if there is no other ε -cover with lower cardinality. The cardinality of any minimal ε-cover is the ε -covering number, denoted by $\mathrm{Cov}(\mathcal{A}, \rho, \varepsilon)$.
- For $a \in \mathcal{A}$ where $\mathcal{A}$ is a function class, define the following norms on $\mathcal{A}$:

$$\|a\|_{p,x} := \left(\frac{1}{n}\sum_{i=1}^{n} |a(x_i)|^p\right)^{1/p} \quad \text{for any } p \in [1, \infty),$$
$$\|a\|_{\infty,x} := \max_i |a(x_i)|.$$

- **Property 10** If $1 \leq p \leq q$, then

$$\mathrm{Cov}(\mathcal{A}, \|\cdot\|_{q,x}, \varepsilon) \leq \mathrm{Cov}(\mathcal{A}, \|\cdot\|_{p,x}, \varepsilon).$$

Covering Number: Theorem

- **Theorem 11 (Chaining)**
  For any $x = \{x_1, \ldots, x_m\} \in \mathcal{X}^m$ and $\sup_{a \in \mathcal{A}} \|a\|_{2,x} \leq c_x$ we have

$$\mathrm{R}(\mathcal{A} \circ x) \leq \frac{12}{\sqrt{m}} \int_0^{c_x/2} \mathrm{d}\nu \sqrt{\log \mathrm{Cov}\left(\mathcal{A}, \|\cdot\|_{2,x}, \nu\right)}.$$

Proof of Theorem 11

- Let $\epsilon_j = 2^{-j} c_x$ for $j = 1, \ldots, m$ be successively finer resolutions.
- For each $j = 0, \ldots, m$, let C_j be an ϵ_j -cover of $\mathcal{A}$.
- Fix any $a \in \mathcal{A}$.
- Let $g_j \in C_j$ be such that $\|a - g_j\| \leq \epsilon_j$; take $g_0 = 0$. Note that g_j 's depend on a.
- Let us decompose a as follows:

$$a = a - g_m + \underbrace{g_0}_{=0} + \sum_{j=1}^{m} (g_j - g_{j-1}) .$$

- Let us bound some norms:

  * $\|a - g_m\| \leq \epsilon_m$
  * $\|g_j - g_{j-1}\| \leq \|g_j - a\| + \|a - g_{j-1}\| \leq \epsilon_j + \epsilon_{j-1} = 3\epsilon_j$ (since $2\epsilon_j = \epsilon_{j-1}$)

Proof of Theorem 11

$$
\begin{aligned}
\mathrm{R}(\mathcal{A}) &= \mathbb{E}\left[\sup_{a\in\mathcal{A}}\langle\sigma, a\rangle\right] && [\text{ definition }]\\
&= \mathbb{E}\left[\sup_{a\in\mathcal{A}}\langle\sigma, a-g_m\rangle + \sum_{j=1}^{m}\langle\sigma, g_j - g_{j-1}\rangle\right] && [\text{ decompose } a]\\
&\leq \epsilon_m + \mathbb{E}\left[\sup_{a\in\mathcal{A}}\sum_{j=1}^{m}\langle\sigma, g_j - g_{j-1}\rangle\right] && [\text{ Cauchy-Schwartz }]\\
&\leq \epsilon_m + \sum_{j=1}^{m}\mathbb{E}\left[\sup_{a\in\mathcal{A}}\langle\sigma, g_j - g_{j-1}\rangle\right] && [\text{ push sup inside }]\\
&\leq \epsilon_m + \sum_{j=1}^{m}\mathbb{E}\left[\sup_{g_j\in C_j, g_{j-1}\in C_{j-1}}\langle\sigma, g_j - g_{j-1}\rangle\right] && [\text{ refine dependence }]\\
&\leq (\textit{next page})
\end{aligned}
$$

Proof of Theorem 11

$$\leq \epsilon_m + \sum_{j=1}^{m} (3\epsilon_j) \sqrt{\frac{2 \log \left(|C_j| \, |C_{j-1}| \right)}{m}} \quad \text{[Massart's lemma (Property 6)]}$$

$$\leq \epsilon_m + \sum_{j=1}^{m} (6\epsilon_j) \sqrt{\frac{\log |C_j|}{m}} \quad [\text{ since } |C_j| \geq |C_{j-1}|]$$

$$= \epsilon_m + \sum_{j=1}^{m} 12 \left(\epsilon_j - \epsilon_{j+1} \right) \sqrt{\frac{\log |C_j|}{m}} \quad [\text{ since } \epsilon_j = 2 \left(\epsilon_j - \epsilon_{j+1} \right)$$

$$\leq 12 \int_0^{c_x/2} \sqrt{\frac{\log \operatorname{Cov} \left(\mathcal{A}, \| \cdot \|_{2,x}, \nu \right)}{m}} d\epsilon \quad [\text{ bound sum with integral }]$$

Covering Number: Examples

- Let $\mathcal{A}_\infty := \left\{x \in \mathbb{R}^d \to w^\top x : w \in \mathbb{R}^d, \|w\|_\infty \leq 1\right\}$. Then, for any $x = \{x_1, \ldots, x_n\}$, please prove

$$\mathrm{R}\left(\mathcal{A}_\infty \circ x\right) \leq 12\gamma \frac{\max_i \|x_i\|_1}{\sqrt{n}} \sqrt{d},$$

  where $\gamma := \int_0^{1/2} \mathrm{d}\nu \sqrt{\log(3/\nu)}$.

Covering Number: Examples

Proof:

- With **Theorem 11** and **Property 10**, we have

$$\mathrm{R}\left(\mathcal{A}_\infty \circ x\right) \leq \frac{12}{\sqrt{n}} \int_0^{c_x/2} \mathrm{d}\nu \sqrt{\log \mathrm{Cov}\left(\mathcal{A}_\infty, \|\cdot\|_{\infty,x}, \nu\right)}.$$

- By Holder's inequality, $a(x) = w^\top x \leq \|w\|_\infty \|x\|_1 \leq \|x\|_1$, so

$$c_x = \sup_{a \in \mathcal{A}_\infty} \|a\|_{2,x} = \sup_{a \in \mathcal{A}_\infty} \sqrt{\frac{1}{n} \sum_{i=1}^n a\left(x_i\right)^2} \leq \sqrt{\frac{1}{n} \sum_{i=1}^n \|x_i\|_1^2} \leq c$$

with $c := \max_i \|x_i\|_1$.

Covering Number: Examples

Proof (cont'd):

- Now let's calculate covering number. Also by Holder's, we get

$$\|a-b\|_{\infty,x} = \max_i |a(x_i) - b(x_i)| = \max_i \left|(w_a - w_b)^\top x_i\right| \leq c\,\|w_a - w_b\|_\infty ;$$

  hence, to find an ν -cover, it suffices to find a finite set $\mathcal{C}$ such that for any w_a, there exists $w \in \mathcal{C}$ with $\|w_a - w\|_\infty \leq \nu/c$.
- For a hypercube with side length 2, divide it into small cubes with side length $2\nu/c$.
- Define $\mathcal{C}$ to be the set of vertices of the cubes.
- Then, any $w_a \in \left\{w \in \mathbb{R}^d : \|w\|_\infty \leq 1\right\}$ must land in one of these cubes, and each coordinate is at most ν/c away from one of the vertices.

Covering Number: Examples

Proof (cont'd):

- There are at most $(\lceil c/\nu \rceil + 1)^d$ vertices, so

$$\mathrm{Cov}\left(\mathcal{A}_\infty, \|\cdot\|_{\infty,x}, \nu\right) \leq (\lceil c/\nu \rceil + 1)^d \leq (c/\nu + 2)^d \leq (3c/\nu)^d.$$

- Finally,

$$\begin{aligned}\mathrm{R}\left(\mathcal{A}_\infty \circ x\right) &\leq \frac{12\sqrt{d}}{\sqrt{n}} \int_0^{c/2} \mathrm{d}\nu \sqrt{\log(3c/\nu)} \\ &= \frac{12c\sqrt{d}}{\sqrt{n}} \int_0^{1/2} \mathrm{d}\nu \sqrt{\log(3/\nu)}.\end{aligned}$$

Reference

1 Wainwright, Martin J. High-dimensional statistics: A non-asymptotic viewpoint. Vol. 48. Cambridge University Press, 2019.

2 Shalev-Shwartz, Shai, and Shai Ben-David. Understanding machine learning: From theory to algorithms. Cambridge university press, 2014.

3 Ji, Ziwei, and Matus Telgarsky. "Polylogarithmic width suffices for gradient descent to achieve arbitrarily small test error with shallow ReLU networks." arXiv preprint arXiv:1909.12292 (2019).

4 Patrick Rebeschini. Lecture notes for the course Algorithmic Foundations of Learning. http://www.stats.ox.ac.uk/~rebeschi/teaching/AFoL/19/.

5 Percy Liang. Lecture notes for the course Statistical Learning Theory.