Rademacher Complexity and Concentration Inequalities

Wei Xiong, Yong Lin

Department of Mathematics, Department of Computer Science Engineering
The Hong Kong University of Science and Technology

2022.3.24

1 Review of ERM and Uniform Convergence by Covering Number

Objectives: derive generalization bound for certain problems

- Empirical Risk Minimization: given n i.i.d. samples in $\mathcal{S}_n$, we take

$$\hat{f} = \operatorname*{argmin}_{f \in \mathcal{F}} \hat{L}(f) := \operatorname*{argmin}_{f \in \mathcal{F}} \frac{1}{n} \sum_{i=1}^{n} \ell(f(X_i), y_i).$$

- Let $L(f)$ be the population loss $\mathbb{E}_{X,y}\ell(f(X), y)$. Then,

$$\begin{aligned} L(\hat{f}) - L(f^*) &= \underbrace{\left(L(\hat{f}) - \hat{L}(\hat{f})\right)}_{A} + \underbrace{\left(\hat{L}(\hat{f}) - \hat{L}(f^*)\right)}_{B} + \underbrace{\left(\hat{L}(f^*) - L(f^*)\right)}_{C} \\ &\leq \left(L(\hat{f}) - \hat{L}(\hat{f})\right) + \left(\hat{L}(f^*) - L(f^*)\right) \\ &\leq 2\sup_{f \in \mathcal{F}} |L(f) - \hat{L}(f)|. \end{aligned}$$

- We cannot apply concentration inequality for $L(\hat{f}) - \hat{L}(\hat{f})$ since $\hat{f}$ also depends on the dataset!
- Solution: uniform convergence!

Uniform Convergence Implies Generalization

- *Finite* $\mathcal{F}$.

$$\begin{aligned}
&P\left(\sup_{f\in\mathcal{F}}|L(f)-\hat{L}(f)|>\sqrt{\frac{1}{2n}\log\frac{2|\mathcal{F}|}{\delta}}\right)\\
&\le\sum_{f\in\mathcal{F}}P\left(|L(f)-\hat{L}(f)|>\sqrt{\frac{1}{2n}\log\frac{2|\mathcal{F}|}{\delta}}\right)\\
&\le|\mathcal{F}|\times\frac{\delta}{|\mathcal{F}|}=\delta,
\end{aligned}\tag{0.1}$$

- *Infinite* $\mathcal{F}$: let $N_\infty(\mathcal{F},\epsilon)$ be the covering number,

$$\begin{aligned}
|L(f)-\hat{L}(f)|&=|\frac{1}{n}\sum_{i=1}^{n}((f_\epsilon-\mathbb{E}f_\epsilon)+(f-f_\epsilon)+\mathbb{E}(f_\epsilon-f))(X_i)|\\
&\le|\frac{1}{n}\sum_{i=1}^{n}(f_\epsilon-\mathbb{E}f_\epsilon)|+|\frac{1}{n}\sum_{i=1}^{n}(f-f_\epsilon)|+|\frac{1}{n}\sum_{i=1}^{n}\mathbb{E}(f_\epsilon-f)|\\
&\le\sqrt{\frac{1}{2n}\log\frac{2N_\infty(\mathcal{F},\epsilon)}{\delta}}+2\epsilon,
\end{aligned}\tag{0.2}$$

Uniform Convergence Implies Generalization

Summary so far.

- ERM:

$$L(\hat{f}) - L(f^*) \leq 2\sup_{f\in\mathcal{F}} |L(f) - \hat{L}(f)|.$$

- Roughly speaking, we are concerning a *uniform* convergence for all $f \in \mathcal{F}$ instead of fixed f where traditional LLN applies;
- What we have learned: uniform convergence via *Covering Number*;

A different but highly related method: Rademacher complexity

- We will first define the Rademacher complexity;
- We show RC is *Sufficient* and *Necessary* for generalization;
- We then estimate RC through various ways.

Remarks

- ERM:

$$\min_{w \in \mathbb{R}^d} \phi(w, \mathcal{S}_n); \qquad \text{No regularization}$$

$$\min_{w \in \mathbb{R}^d} \phi(w, \mathcal{S}_n) + h(w); \qquad \text{Data-independent regularization, e.g. L1,L2}$$

$$\min_{w \in \mathbb{R}^d} \phi(w, \mathcal{S}_n) + h(w, \mathcal{S}_n) \qquad \text{Data-dependent regularization}$$

- Lots of important lemmas from chapter 4 are unavailable;
- Missing contents are included and only the first ERM formulation is considered for simplicity;
- Also, $f \in \mathcal{F}$ is used instead of $w \in \Omega$. We can take

$$\mathcal{F} = \{w \in \Omega : \phi(w, \mathcal{S}_n)\}.$$

- We assume $\hat{f} \in \mathcal{F}$ is the solution of ERM.

Examples from Statistics

- The estimation of CDF:

$$\hat{F}_n(t) = \frac{1}{n}\sum_{i=1}^{n} I(X_i \leq t);$$

  - For *each* $t \in \mathbb{R}$, $\hat{F}_n(t) \to F(t)$ a.s. by SLLN;
  - *Glivenko-Cantelli*: we have the uniform convergence result:

$$\left\| \hat{F}_n - F \right\|_\infty \to 0, a.s..$$

- More general, we consider a function class $\mathcal{F}$ and consider

$$\|\mathbb{P}_n - \mathbb{P}\|_{\mathcal{F}} := \sup_{f\in\mathcal{F}} \left| \frac{1}{n}\sum_{i=1}^{n} f(X_i) - \mathbb{E}f(X) \right|.$$

  - Classical Glivenko-Cantelli corresponds to $\mathcal{F} = \{I(x \leq t) : t \in \mathbb{R}\}$.

Example: Uniform Convergence Can Fail

Let $\mathcal{S} = \{S \subset [0,1] : |S| < \infty\}$ and let $\mathcal{F}_{\mathcal{S}} = \{I_S(\cdot) : S \in \mathcal{S}\}$ be the set of indicator functions. Suppose $X_i \sim U([0,1])$ with $P(\{x\}) = 0$ for all $x \in [0,1]$.

- It holds that $P(S) \leq \sum_{s \in S} P(s) = 0$ for all $S \in \mathcal{S}$;
- In particular, $P(\{X_1, \cdots, X_n\}) = 0, \quad \forall n \in N$;
- By the definition of $\mathbb{P}_n(\cdots)$ (or $\hat{F}_n$), we have $P_n(\{X_1, \cdots, X_n\}) = 1$;
- Then,

$$\sup_{S \in \mathcal{S}} |\mathbb{P}_n(S) - \mathbb{P}(S)| = 1 - 0 = 1$$

2 Rademacher Complexity

Rademacher Complexity

Definition 1

Let $\sigma = (\sigma_1, \cdots, \sigma_n)$ be iid ± 1 Bernoullis r.v.s with $p = 1/2$.

- Empirical Rademacher complexity: let $S = (x_1, \cdots, x_n)$:

$$R_S(\mathcal{F}) := \mathbb{E}_\sigma \sup_{f \in \mathcal{F}} \left| \frac{1}{n} \sum_{i=1}^{n} \sigma_i f(x_i) \right|$$

- Rademacher Compelxity:

$$R_n(\mathcal{F}) = \mathbb{E}_S R_S(\mathcal{F}) = \mathbb{E}_{X,\sigma} \sup_{f \in \mathcal{F}} \left| \frac{1}{n} \sum_{i=1}^{n} \sigma_i f(x_i) \right|.$$

RC is Sufficient for Uniform Convergence

Theorem 2

We consider b-uniformly bounded $\mathcal{F}$, i.e., $\|f\|_\infty \leq b, \forall f \in \mathcal{F}$. Then, $\forall t > 0$, w.p. at least $1-\delta$.

$$\|\mathbb{P}_n - \mathbb{P}\|_{\mathcal{F}} = \sup_{f \in \mathcal{F}} \left| \frac{1}{n} \sum_{i=1}^{n} f(X_i) - \mathbb{E}f(X) \right| \leq 2R_n(\mathcal{F}) + \sqrt{\frac{2b \log 1/\delta}{n}};$$

The bound in expectation:

$$\mathbb{E}_{X^n} \|\mathbb{P}_n - \mathbb{P}\|_{\mathcal{F}} \leq 2R_n(\mathcal{F})$$

- $R_n(\mathcal{F}) = o(1)$ is sufficient for $\|\mathbb{P}_n - \mathbb{P}\|_{\mathcal{F}} \to 0$ almost surely because

$$\sum_{n=1}^{\infty} P(\|\mathbb{P}_n - \mathbb{P}\|_{\mathcal{F}} > 2R_n(\mathcal{F}) + \delta) \leq \sum_{n=1}^{\infty} \exp(-\frac{n\delta^2}{2b^2}) < \infty.$$

- BC Lemma implies that $P(\limsup_n \{\|\mathbb{P}_n - \mathbb{P}\|_{\mathcal{F}} > 2R_n(\mathcal{F}) + \delta\}) = 0$.

Proof of Bound in Expectation: $\mathbb{E}_{X^n} \|\mathbb{P}_n - \mathbb{P}\|_{\mathcal{F}} \leq 2R_n(\mathcal{F})$

Part I.

We first sample $X_1', \cdots, X_n'$ from $\mathbb{P}$ which are independent with X^n. Then,

$$\begin{aligned}
\sup_{f\in\mathcal{F}} \left(\frac{1}{n}\sum_{i=1}^{n} f(X_i) - \mathbb{E}[f]\right) &= \sup_{f\in\mathcal{F}} \left(\frac{1}{n}\sum_{i=1}^{n} f(X_i) - \underset{X_1',\ldots,X_n'}{\mathbb{E}} \left[\frac{1}{n}\sum_{i=1}^{n} f(X_i')\right]\right) \\
&= \sup_{f\in\mathcal{F}} \left(\mathrm{E}_{X_1',\ldots,X_n'} \left[\frac{1}{n}\sum_{i=1}^{n} f(X_i) - \frac{1}{n}\sum_{i=1}^{n} f(X_i')\right]\right) \\
&\leq \mathbb{E}_{X_1',\ldots,X_n'} \left[\sup_{f\in\mathcal{F}} \left(\frac{1}{n}\sum_{i=1}^{n} f(X_i) - \frac{1}{n}\sum_{i=1}^{n} f(X_i')\right)\right]
\end{aligned}$$

where the last step uses

$$\sup_{u} \mathbb{E}_v g(u,v) \leq \sup_{u} \mathbb{E}_v \sup_{u'} g(u',v) = \mathbb{E}_v \sup_{u} g(u,v).$$

$\square$

Proof of Bound in Expectation: $\mathbb{E}_{X^n} \|\mathbb{P}_n - \mathbb{P}\|_{\mathcal{F}} \leq 2R_n(\mathcal{F})$

Part II.

We have

$$\sup_{f\in\mathcal{F}} \left(\frac{1}{n}\sum_{i=1}^{n} f\left(X_i\right) - \mathbb{E}[f]\right) \leq \mathbb{E}_{X_1',\ldots,X_n'}\left[\sup_{f\in\mathcal{F}} \left(\frac{1}{n}\sum_{i=1}^{n} f\left(X_i\right) - \frac{1}{n}\sum_{i=1}^{n} f\left(X_i'\right)\right)\right]$$

We then take expectation over X^n to obtain

$$\begin{aligned}
&\mathbb{E}_{X^n} \sup_{f\in\mathcal{F}} \left(\frac{1}{n}\sum_{i=1}^{n} f\left(X_i\right) - \mathbb{E}[f]\right) \leq \mathbb{E}_{X^n}\mathbb{E}_{(X')^n} \sup_{f\in\mathcal{F}} \left(\frac{1}{n}\sum_{i=1}^{n} f\left(X_i\right) - \frac{1}{n}\sum_{i=1}^{n} f\left(X_i'\right)\right) \\
&= \mathbb{E}_{X^n}\mathbb{E}_{(X')^n}\mathbb{E}_{\sigma} \sup_{f\in\mathcal{F}} \left(\frac{1}{n}\sum_{i=1}^{n} \sigma_i\left(f(X_i) - f\left(X_i'\right)\right)\right) \\
&\leq \mathbb{E}_{X^n(X')^n\sigma}\left[\sup_{f}\left(\frac{1}{n}\sum_{i=1}^{n}\sigma_i f(X_i)\right) + \sup_{f}\left(\frac{1}{n}\sum_{i=1}^{n}\sigma_i f(X_i')\right)\right] = 2R_n(\mathcal{F}).
\end{aligned}$$

because $\sigma_i(f(X_i) - f(X_i'))$ has the same distribution with $f(X_i) - f(X_i')$. □

Proof of High-Probability Bound

We need the following concentration inequality to prove w.p. $1-\delta$, $\|\mathbb{P}_n - \mathbb{P}\|_{\mathcal{F}} = \sup_{f\in\mathcal{F}} \left|\frac{1}{n}\sum_{i=1}^n f(X_i) - \mathbb{E}f(X)\right| \le 2R_n(\mathcal{F}) + \sqrt{\frac{2b\log 1/\delta}{n}}$.

Lemma 3 (McDiarmid's inequality)

Suppose $f : \mathbb{R}^n \to \mathbb{R}$ satisfies that for all $x_1, \cdots, x_n, x_i'$, we have

$$|f(x_1, \cdots, x_n) - f(x_1, \cdots, x_{i-1}, x_i', x_{i+1}, \cdots, x_n)| \le c_i$$

for some constants $c_1, \cdots, c_n$. Then,

$$P(f(X_1, \cdots, X_n) - \mathbb{E}f(X_1, \cdots, X_n) \ge t) \le \exp(-\frac{2t^2}{\sum_{i=1}^n c_i^2}).$$

Idea: $\|\mathbb{P}_n - \mathbb{P}\|_{\mathcal{F}}$ is close to $\mathbb{E}_{X^n}\|\mathbb{P}_n - \mathbb{P}\|_{\mathcal{F}}$ with high probability according to McDiarmid's inequality.

Proof of High-Probability Bound

Part I: $\|\mathbb{P}_n - \mathbb{P}\|_{\mathcal{F}}$ concentrates on its mean.

Let $\bar{f}(x) = f(x) - \mathbb{E}f(X)$ and $G(x^n) = \sup_f |\frac{1}{n}\sum_{i=1}^n \bar{f}(x_i)|$. Let $x_i = y_i$ except for $i \neq 1$.

$$|\frac{1}{n}\sum_{i=1}^n \bar{f}(x_i)| - \sup_{h\in\mathcal{F}} |\frac{1}{n}\sum_{i=1}^n \bar{h}(y_i)| \leq |\frac{1}{n}\sum_{i=1}^n \bar{f}(x_i)| - |\frac{1}{n}\sum_{i=1}^n \bar{f}(y_i)| \leq \frac{1}{n}|\bar{f}(x_1) - \bar{f}(y_1)| \leq \frac{2b}{n}.$$

We take supremum over f to obtain $G(x) - G(y) \leq \frac{2b}{n}$ and $G(y) - G(x) \leq \frac{2b}{n}$ similarly. Therefore, $c_i = \frac{2b}{n}$ so

$$\|\mathbb{P}_n - \mathbb{P}\|_{\mathcal{F}} - \mathbb{E}\,\|\mathbb{P}_n - \mathbb{P}\|_{\mathcal{F}} \leq t$$

w.p. at least $1 - \exp(-\frac{nt^2}{2b^2})$ by the McDiarmid's inequality. To be clear, we have

$$\sup_{f\in\mathcal{F}} \left|\frac{1}{n}\sum_{i=1}^n f(X_i) - \mathbb{E}_{X_i} f(X_i)\right| \leq \mathbb{E}_X \sup_{f\in\mathcal{F}} \left|\frac{1}{n}\sum_{i=1}^n f(X_i) - \mathbb{E}_{X_i} f(X_i)\right| + t.$$

Proof of High-Probability Bound

We prove $\sup_{f\in\mathcal{F}} \left|\frac{1}{n}\sum_{i=1}^n f(X_i) - \mathbb{E}f(X)\right| \le 2R_n(\mathcal{F}) + t$ w.p. at least $1-\exp(-\frac{nt^2}{2b^2})$. We denote Y_1^n to be a second i.i.d. sequence independent of X_1^n.

Part II: Combine concentration with symmetric technique.

Recall from the previous slide:

$$\mathbb{E}_X \sup_{f\in\mathcal{F}} \left|\frac{1}{n}\sum_{i=1}^n f(X_i) - \mathbb{E}_{X_i} f(X_i)\right| \le 2R_n(\mathcal{F});$$

$$\sup_{f\in\mathcal{F}} \left|\frac{1}{n}\sum_{i=1}^n f(X_i) - \mathbb{E}_{X_i} f(X_i)\right| \le \mathbb{E}_X \sup_{f\in\mathcal{F}} \left|\frac{1}{n}\sum_{i=1}^n f(X_i) - \mathbb{E}_{X_i} f(X_i)\right| + t.$$

This implies

$$\sup_{f\in\mathcal{F}} \left|\frac{1}{n}\sum_{i=1}^n f(X_i) - \mathbb{E}_{X_i} f(X_i)\right| \le 2R_n(\mathcal{F}) + t$$

w.p. at least $1-\exp(-\frac{nt^2}{2b^2})$. □

RC is Necessary for Uniform Convergence

With $\|\mathcal{S}_n\|_{\mathcal{F}} = \sup_{f\in\mathcal{F}} \left|\frac{1}{n}\sum_{i=1}^n \sigma_i f(X_i)\right|$ and $\bar{\mathcal{F}} = \{f \in \mathcal{F} : f - \mathbb{E}f\}$.

- We have

$$\frac{1}{2}\mathbb{E}_{X,\sigma} \|\mathcal{S}_n\|_{\bar{\mathcal{F}}} \leq \mathbb{E}_X \left[\|\mathbb{P}_n - \mathbb{P}\|_{\mathcal{F}}\right] \leq 2\mathbb{E}_{X,\sigma} \|\mathcal{S}_n\|_{\mathcal{F}}, \qquad (0.3)$$

- The proof is omitted for simplicity. We focus on the following theorem:

Theorem 4

For b-uniformly bounded $\mathcal{F}$, w.p. at least $1-\delta$, we have

$$\|\mathbb{P}_n - \mathbb{P}\|_{\mathcal{F}} \geq \frac{1}{2}R_n(\mathcal{F}) - \frac{\sup_{f\in\mathcal{F}} |\mathbb{E}[f]|}{2\sqrt{n}} - \sqrt{\frac{2b\log(1/\delta)}{n}}. \qquad (0.4)$$

- If the RC of $\bar{F} = w(1)$, the $\|\mathbb{P}_n - \mathbb{P}\|_{\mathcal{F}}$ cannot converge to zero in probability.

Proof of Necessity of Rademacher Complexity

We assume that $\frac{1}{2}\mathbb{E}_{X,\sigma}\left\|\mathcal{S}_n\right\|_{\bar{\mathcal{F}}} \leq \mathbb{E}_X\left[\|\mathbb{P}_n-\mathbb{P}\|_{\mathcal{F}}\right]$. It suffices to lower bound $\frac{1}{2}\mathbb{E}_{X,\sigma}\left\|\mathcal{S}_n\right\|_{\bar{\mathcal{F}}}$.

Proof.

$$\left\|\mathcal{S}_n\right\|_{\overline{\mathcal{F}}}=\sup_{f\in\mathcal{F}}\left|\frac{1}{n}\sum_{i=1}^n\sigma_i f\left(X_i\right)-\frac{1}{n}\sum_{i=1}^n\sigma_i\mathbb{E}[f]\right| \geq \sup_{f\in\mathcal{F}}\left|\frac{1}{n}\sum_{i=1}^n\sigma_i f\left(X_i\right)\right|-\frac{\left|\sum_{i=1}^n\sigma_i\right|\sup_{f\in\mathcal{F}}|\mathbb{E}[f]|}{n}$$

Note $\mathbb{E}\left|\sum_{i=1}^n\sigma_i\right| \leq \sqrt{\mathbb{E}(\sum_{i=1}^n\sigma_i)^2}=\sqrt{\mathbb{E}\sum_{i=1}^n\sigma_i^2}=\sqrt{n}$. Taking expectation, we have

$$\frac{1}{2}\mathbb{E}_{X,\sigma}\left\|\mathcal{S}_n\right\|_{\bar{\mathcal{F}}} \geq \frac{1}{2}R_n(\mathcal{F})-\frac{\sup_{f\in\mathcal{F}}|\mathbb{E}[f]|}{2\sqrt{n}}.$$

It remains note that $\sup_{f\in\mathcal{F}}\left|\frac{1}{n}\sum_{i=1}^n\sigma_i f\left(X_i\right)\right|$ concentrates around its mean $R_n(\mathcal{F})$ by McDiarmid's inequality. □

3 Bounding Rademacher Complexity

RC Can Be Independent with $\mathcal{F}$

- Let $x = x_0$ w.p. 1 and $f(x_0) \in [-1, 1]$ for all $f \in \mathcal{F}$.

$$\mathbb{E}_{X,\sigma}\left[\sup_{f\in\mathcal{F}} \frac{1}{n}\sum_{i=1}^{n}\sigma_i f\left(x_i\right)\right] = \mathbb{E}_{\sigma}\left[\sup_{f\in\mathcal{F}} \frac{1}{n} f\left(x_0\right)\sum_{i=1}^{n}\sigma_i\right] \leq \mathbb{E}_{\sigma}\left[\left|\frac{1}{n}\sum_{i=1}^{n}\sigma_i\right|\right]$$
$$\leq \left[\mathbb{E}_{\sigma}\left(\frac{1}{n}\sum_{i=1}^{n}\sigma_i\right)^2\right]^{\frac{1}{2}} = \frac{1}{n}\left(\mathop{\mathbb{E}}_{\sigma_i,\sigma_j}\left[\sum_{i,j=1}^{n}\sigma_i\sigma_j\right]\right)^{\frac{1}{2}} = \frac{1}{n}\left(\mathop{\mathbb{E}}_{\sigma_i}\left[\sum_{i=1}^{n}\sigma_i^2\right]\right)^{\frac{1}{2}}$$
$$= \frac{1}{n}\cdot\sqrt{n} = \frac{1}{\sqrt{n}}.$$

- We only use the boundedness of f in the second step;
- Natural because $P(x)$ is very easy to learn.

Connection with Covering Method.

- Finite $\mathcal{F}$: let $\sqrt{\frac{1}{n}\sum_{i=1}^{n} f(z_i)^2} \leq M$:

$$R_S(\mathcal{F}) \leq \sqrt{\frac{2M^2 \log |\mathcal{F}|}{n}}.$$

- Infinite 2-uniformly bounded $\mathcal{F}$:

$$R_S(\mathcal{F}) \leq \inf_{\epsilon>0} \left(\epsilon + \sqrt{\frac{2\log(N(\epsilon, \mathcal{F}, L_2(P_n))}{n}} \right),$$

where $L_2(P_n)(f, f') := \sqrt{\frac{1}{n}\sum_{i=1}^{n}(f(x_i) - f'(x_i))^2}$;

- Infinite $\mathcal{F}$, Dudley's Theorem:

$$R_S(\mathcal{F}) \leq 4\alpha + 12\int_{\alpha}^{\infty} \sqrt{\frac{\log N(\epsilon, \mathcal{F}, L_2(P_n))}{n}} d\epsilon; \qquad (0.5)$$

Connection with Covering Method.

We assume $\mathcal{F}$ is 2-uniformly bounded.

- $N(\epsilon, \mathcal{F}, L_2(P_n)) \approx (1/\epsilon)^R$:

$$R_S(\mathcal{F}) \leq c \int_0^1 \sqrt{\frac{R \log(1/\epsilon)}{n}} d\epsilon \approx \sqrt{\frac{R}{n}}.$$

- $N(\epsilon, \mathcal{F}, L_2(P_n)) \approx \exp(R/\epsilon)$:

$$R_S(\mathcal{F}) \leq c \int_0^1 \sqrt{\frac{R/\epsilon}{n}} d\epsilon \approx \sqrt{\frac{R}{n}}.$$

- $N(\epsilon, \mathcal{F}, L_2(P_n)) \approx \exp(R/\epsilon^2)$: with $\alpha = 1/poly(n)$:

$$R_S(\mathcal{F}) \leq 4\alpha + \int_\alpha^1 \sqrt{\frac{R/\epsilon^2}{n}} d\epsilon = \frac{1}{poly(n)} + \sqrt{\frac{R}{n}} \log(1/\alpha) = \tilde{O}(\sqrt{\frac{R}{n}}).$$

4 Local Rademacher Complexity

Comparison of Different Complexities

- Distribution-free methods: Vapnik-Chervonenkis dimension or metric entropy , typically give conservative estimates.
- Distribution-dependent methods: e.g., based on entropy numbers in the $L_2(P)$ distance, are not useful when the underlying distribution P is unknown.
- Data-dependent methods: e.g. Rademacher complexity, can be directly computely from the data,
  - They provide global estimates of the complexity of the function class.
  - They do not reflect the fact that the algorithm will likely pick functions that have a small error.
  - Best rate is $1/\sqrt{n}$.

Preliminary Results

Theorem 5

Suppose $f(X) \in [a, b], \forall f \in \mathcal{F}, X \in \mathcal{X}$. Assume that there is some $r > 0$ such that for every $f \in \mathcal{F}$, $\mathbb{V}[f(X)] \leq r$. Then for any $\delta > 0$, with probability at least $1 - \delta$,

$$\|\mathbb{P}_n - \mathbb{P}\|_{\mathcal{F}} \leq \inf_{\alpha>0} \left(2(1+\alpha)R_n(\mathcal{F}) + \sqrt{\frac{2r \ln 1/\delta}{n}} + (b-a)(\frac{1}{3} + \frac{1}{\alpha})\frac{\ln(1/\delta)}{n}\right)$$

and with probability at least $1 - 2\delta$,

$$\|\mathbb{P}_n - \mathbb{P}\|_{\mathcal{F}} \leq$$
$$\inf_{\alpha>0} \left(2\frac{1+\alpha}{1-\alpha}R_s(\mathcal{F}) + \sqrt{\frac{2r \ln(1/\delta)}{n}} + (b-a)(\frac{1}{3} + \frac{1}{\alpha} + \frac{1+\alpha}{2\alpha(1-\alpha)})\frac{\ln(1/\delta)}{n}\right)$$

Preliminary Results

- When applied to full function space $\mathcal{F}$, Theorem 5 is not useful, which results in $\sqrt{r \ln(1/\delta)/n}$. This is even inferior to $\sqrt{\ln(1/\delta)/n}$ obtained by bounded difference inequality.
- It is meaningful to apply Thereorm 5 to a subset of $\mathcal{F}$.

Preliminary Results

Let $R_n(f) := \frac{1}{n}\sum_{i=1}^{n} \sigma_i f(X_i)$, $R_n(\mathcal{F}) = \sup_{f\in\mathcal{F}} R_n(f)$, $\mathbb{E}[f] := \mathbb{E}_X[f(X)]$ and $\hat{\mathbb{E}}[f^2] := \frac{1}{n}\sum_{i=1}^{n} f^2(x_i)$.

Theorem 6

If $\forall x \in \mathcal{X}, f \in \mathcal{F}, f(x) \in [-b, b]$, then for every $\delta > 0$ and r that satisfy

$$r \geq 10bR_n(\{f : f \in \mathcal{F}, \mathbb{E}[f^2] \leq r\}) + \frac{11b^2 \ln(1/\delta)}{n}$$

then with probability at least $1 - \delta$,

$$\{f \in \mathcal{F} : \mathbb{E}[f^2] \leq r\} \subset \{f \in \mathcal{F} : \hat{\mathbb{E}}[f^2] \leq 2r\}.$$

Preliminary Results

Proof.

Since the range of the function in the set $\mathcal{F}_r = \{f^2 : f \in \mathcal{F}, \mathbb{E}f^2 \leq r\}$ is contained in $[0, b^2]$, it follows that $\mathbb{V}[f^2(X_i)] \leq \mathbb{E}f^4 \leq b^2 Ef^2 \leq b^2 r$. Then by applying Theorem 5 (with $\alpha = 1/4$), with probability at least $1 - \delta$, every $f \in \mathcal{F}_r$ satisfies,

$$
\begin{aligned}
\hat{\mathbb{E}}f^2 &\leq \mathbb{E}f^2 + \frac{5}{2}R_n(\mathcal{F}_r) + \sqrt{\frac{2b^2 r \ln(1/\delta)}{n}} + \frac{13b^2 \ln(1/\delta)}{3n} \\
&\leq r + \frac{5}{2}R_n(\mathcal{F}_r) + \frac{r}{2} + \frac{16b^2 \ln(1/\delta)}{3n} \\
&\leq r + 5bR_n(\{f : f \in \mathcal{F}, \mathbb{E}f^2 \leq r\}) + \frac{r}{2} + \frac{16b^2 \ln(1/\delta)}{3n} \\
&\leq 2r
\end{aligned}
$$

where the last but one inequality is applying the Lipschitz function Theorem 6.6 in the lecture note with $\phi(x) = x^2$ and Lipschitz constant $2b$. □

Error bounds with local complexity

Definition 7

A function $\psi : [0, \infty) \to [0, \infty)$ is sub-root if it is nonnegative, nondecreasing and if $r \to \psi(r)/\sqrt{r}$ is nonincreasing for $r > 0$.

Lemma 8

If $\psi : [0, \infty) \to [0, \infty)$ is a nontrivial sub-root function, then it is continuous on $[0, \infty)$ and the equation $\psi(r) = r$ has a unique positive solution. Moreover, if we denote the solution by r^*, then for all $r > 0$, $r \geq \psi(r)$ if and only if $r* \leq r$.

Error bounds with local complexity

Theorem 9

Let $\mathcal{F}$ be a class of functions with ranges in $[a, b]$ and assume that there are some functional $T : \mathcal{F} \to \mathbb{R}^+$ and some constant B such that for every $f \in \mathcal{F}$, $\mathbb{V}[f] \leq T(f) \leq BPf$. Let ψ be a sub-root function and let r^* be the fixed point of ψ. Assume that ψ satisfies, for any $r \geq r^*$,

$$\psi(r) \geq BR_n(\{f \in \mathcal{F} : T(f) \leq r\}).$$

Then, with $c_1 = 704$ and $c_2 = 26$, for any $K > 1$ and every $\delta > 0$, with probability at least $1 - \delta$,

$$\forall f \in \mathcal{F}, \mathbb{E}f \leq \frac{K}{K-1}\hat{\mathbb{E}}f + \frac{c_1 K}{B} r^* + \frac{\ln(1/\delta)(11(b-a) + c_2 BK)}{n}.$$

Moreover, for $f \in \mathcal{F}$ and $\alpha \in [0, 1]$, $T(\alpha f) \leq \alpha^2 T(f)$, and if ψ satisfies, for any $r \geq r^*$, $\psi(r) \geq BR_n(\{f \in star(\mathcal{F}, 0) : T(f) \leq r\})$, then the same results hold true with $c_1 = 6$ and $c_2 = 5$.

Error bounds with local complexity

The main technique:

- *peeling*, partition the function class $\mathcal{F}$ into slices where functions have variance within a certain range.
- *re-weighting*, re-weighting the functions in $\mathcal{F}$ by dividing them by their variance.

The proof road map:

- Apply Theorem 5 to the class $\{f/\mathbb{V}[f] : f \in \mathcal{F}\}$. By such re-weighting, the functions have a small variance.
- 'peeling off' subclasses of $\mathcal{F}$ according to the variance of their elements, bounding Rademacher complexity of these subclasses by ψ.
- Using the sub-root property of ψ, so that its fix point gives a common upper of the complexity.
- Convert the bound on the reweighted class to the original class.

Error bounds with local complexity

Given a class $\mathcal{F}$, $\lambda > 1$ and $r > 0$, let
$w(f) = \min\{r\lambda^k : k \in \mathbb{N}, r\lambda^k \geq T(f)\}$ and set $\mathcal{G}_r = \left\{\frac{r}{w(f)}f : f \in \mathcal{F}\right\}$.
Because $w(f) \geq r$, so $\mathcal{G}_r \subset \{\alpha f : f \in \mathcal{F}, \alpha \in [0,1]\} = \mathrm{star}(\mathcal{F}, 0)$.

Lemma 10 (Weighted Function to Original Function)

With the above notation, assume that there is a constant $B > 0$ such that for every $f \in \mathcal{F}$, $T(f) \leq B\mathbb{E}f$. Fix $K > 0$, $\lambda > 0$ and $r > 0$. If $(\mathbb{P} - \mathbb{P}_n)_{\mathcal{G}_r} \leq r/(\lambda BK)$, then

$$\forall f \in \mathcal{F}, \quad \mathbb{E}f \leq \frac{K}{K-1}\hat{\mathbb{E}}f + \frac{r}{\lambda BK}.$$

The condition $(\mathbb{P} - \mathbb{P}_n)_{\mathcal{G}_r} \leq r/(\lambda BK)$ is crucial.

Error bounds with local complexity

Proof.

Notice that for all $g \in \mathcal{G}_r$, $\mathbb{E}g \leq \hat{\mathbb{E}}g + (\mathbb{P} - \mathbb{P}_n)_{\mathcal{G}_r}$. Fix $f \in \mathcal{F}$ and define $g = rf/w(f)$.
Case 1: when $T(f) \leq r$, $w(f) = r$, so that $g = f$. Thus, we have

$$\mathbb{E}f \leq \hat{\mathbb{E}}f + (\mathbb{P} - \mathbb{P}_n)_{\mathcal{G}_r} \leq \hat{\mathbb{E}}f + r/(\lambda BK).$$

Case 2: if $T(f) > r$, then $w(f) = r\lambda^k$ with $k > 0$ and $T(f) \in (r\lambda^{k-1}, r\lambda^k]$. Moreover, $g = f/\lambda^k$, so $\mathbb{E}g \leq \hat{\mathbb{E}}g + (\mathbb{P} - \mathbb{P}_n)_{\mathcal{G}_r}$ leads to the following:

$$\frac{\mathbb{E}f}{\lambda^k} \leq \frac{\hat{\mathbb{E}}f}{\lambda^k} + (\mathbb{P} - \mathbb{P}_n)_{\mathcal{G}_r}$$

Using the fact that $T(f) > r\lambda^{k-1}$, it follows that

$$\mathbb{E}f \leq \hat{\mathbb{E}}f + \lambda^k(\mathbb{P} - \mathbb{P}_n)_{\mathcal{G}_r} \leq \hat{\mathbb{E}}f + \lambda T(f)/r(\mathbb{P} - \mathbb{P}_n)_{\mathcal{G}_r} \leq \hat{\mathbb{E}}f + \frac{r\lambda B\mathbb{E}f}{\lambda BKr}.$$

Error bounds with local complexity

Proof of Theorem 9.

Let $\mathcal{G}_r$ be defined as above, where r is chosen such that $r \geq r^*$, and note that functions in $\mathcal{G}_r$ satisfy $\|g - \mathbb{E}g\|_\infty \leq b - a$ since $0 \leq r/w(f) \leq 1$. Also, we have $\mathbb{V}[f] \leq r$.

- If $T(f) > r$, $g = f/\lambda^k$, where k is such that $T(f) \in (r\lambda^{k-1}, r\lambda^k]$, so that $\mathbb{V}[g] = \mathbb{V}[f]/\lambda^{2k} \leq r$.
- If $T(f) \leq r$, $g = f$.

Applying Theorem 5, for any $\delta > 0$, with probability at least $1 - \delta$,

$$(\mathbb{P} - \mathbb{P}_n)_{\mathcal{G}_r} \leq 2(1+\alpha)\mathbb{E}R_n(\mathcal{G}_r) + \sqrt{\frac{2r\ln 1/\delta}{n}} + (b-a)(\frac{1}{3} + \frac{1}{\alpha})\frac{\ln 1/\delta}{n}$$

□

Error bounds with local complexity

Proof of Theorem 9 (Cont.)

Let $\mathcal{F}(x,y) := \{f \in \mathcal{F} : x \leq T(f) \leq y\}$ and define t to be the smallest integer such that $r\lambda^{t+1} \geq Bb$ (B is a chosen constant, b is the upper bound of the f). Then

$$\begin{aligned}
\mathbb{E}R_n(\mathcal{G}_r) \leq & \mathbb{E}R_n(\mathcal{F}(0,r)) + \mathbb{E}\sup_{f\in\mathcal{F}(r,Bb)} \frac{r}{w(f)} R_n(f) \\
\leq & \mathbb{E}R_n(\mathcal{F}(0,r)) + \sum_{j=0}^{t} \mathbb{E} \sup_{f\in\mathcal{F}(r\lambda^j, r\lambda^{j+1})} \frac{r}{w(f)} R_n(f) \\
= & \mathbb{E}R_n(\mathcal{F}(0,r)) + \sum_{j=0}^{t} \lambda^{-j} \mathbb{E} \sup_{f\in\mathcal{F}(r\lambda^j, r\lambda^{j+1})} R_n(f) \\
\leq & \frac{\psi(r)}{B} + \frac{1}{B}\sum_{j=0}^{t} \lambda^{-j}\psi(r\lambda^{j+1}).
\end{aligned}$$

The last inequality is due to the assumption of $\psi(r)$ in Theorem 9.

Error bounds with local complexity

Proof of Theorem 9 (Cont.)

By our assumption, it follows that for $\beta \geq 1$, $\psi(\beta r) \geq \sqrt{\beta}\psi(r)$. Hence

$$\mathbb{E}R_n(\mathcal{G}_r) \leq \frac{1}{B}\Big(1 + \sqrt{\lambda}\sum_{j=0}^{k}\lambda^{-j/2}\Big)\psi(r).$$

Taking $\lambda = 4$ makes the RHS upper bounded by $5\psi(r)/B$. Moreover, for $r \geq r^*$, $\psi(r) \leq \sqrt{r/r^*}\psi(r^*) = \sqrt{rr^*}$, and thus

$$\begin{aligned}
(\mathbb{P} - \mathbb{P}_n)_{\mathcal{G}_r} &\leq \frac{10(1+\alpha)}{B}\sqrt{rr^*} + \sqrt{\frac{2r\ln(1/\delta)}{n}} + (b-a)(\frac{1}{3} + \frac{1}{\alpha})\frac{x}{n} \\
&= (\frac{10(1+\alpha)}{B}\sqrt{r^*} + \sqrt{\frac{2\ln(1/\delta)}{n}})\sqrt{r} + (b-a)(\frac{1}{3} + \frac{1}{\alpha})\frac{x}{n} \\
&= A\sqrt{r} + C
\end{aligned}$$

$\square$

Proof of Theorem 9 (Cont.)

We want to show $(\mathbb{P}-\mathbb{P}_n)_{\mathcal{G}_r} \leq r/(\lambda BK)$. So we can chose $r = r_0$ where r_0 is the largest sol of $A\sqrt{r} + C = r/(\lambda BK)$. We then have $r = r_0 \leq (\lambda BK)^2 A^2 + 2\lambda BKC$. Applying this to Lemma 10 leads to the following:

$$\begin{aligned}
\mathbb{E}f &\leq \frac{K}{K-1}\hat{\mathbb{E}}f + \frac{r}{\lambda BK} \\
&\leq \frac{K}{K-1}\hat{\mathbb{E}}f + \lambda BKA^2 + 2C \\
&\leq \frac{K}{K-1}\hat{\mathbb{E}}f + \frac{c_1 K}{B} r^* + \frac{\ln(1/\delta)(11(b-a) + c_2 BK)}{n}.
\end{aligned}$$

□

Estimating r^*

As shown in Theorem 9, the error bound involves r^*, which is the fixed point of $\psi(r)$. We need to estimate r^* when applying this result. Let's first choose the function ψ. $star(\mathcal{F}, f_0) = \{f_0 + \alpha(f - f_0) : f \in \mathcal{F}, \alpha \in [0, 1]\}$.

Lemma 11

If the class $\mathcal{F}$ is star-shaped around $\hat{f}$ (which may depend on the data), and $T : \mathcal{F} \to \mathbb{R}^+$ is a function that satisfies $T(\alpha f) \le \alpha^2 T(f)$ for any $f \in \mathcal{F}$ and any $\alpha \in [0, 1]$, then the function ψ defined for $r \ge 0$ by

$$\psi(r) = \mathbb{E}_\sigma R_n(\{f \in \mathcal{F} : T(f - \hat{f}) \le r\})$$

is sub-root and $r \to \mathbb{E}\psi(r)$ is also sub-root.

Notice that making a class star-shaped only increases it, so that

$$\mathbb{E}R_n(f \in star(\mathcal{F}, f_0) : T(f) \le r) \ge \mathbb{E}R_n(f \in \mathcal{F} : T(f) \le r)$$

Estimating r^*

Taking $\hat{f} = 0$ and consider $star(\mathcal{F}, 0)$. We will show $\psi(r) = \mathbb{E}_\sigma R_n(\{f \in star(\mathcal{F}, 0) : T(f) \leq r\})$ is subroot. It is easy to show $\psi(r) \geq 0$ and $\psi(r)$ is non-decreasing. It remains to show for any $0 < r_1 \leq r_2$, $\psi(r_1) \geq \sqrt{r_1/r_2}\psi(r_2)$.

Proof.

Fixing any sample and any realization of the Rademacher random variables, and set f_0 to be a function for which $\sup_{f \in star(\mathcal{F},0), T(f) \leq r_2} \sum_{i=1}^n \sigma_i f(x_i)$ is attained. Since $T(f_0) \leq r_2$, then $T(\sqrt{r_1/r_2} f_0) \leq r_1$ by assumption. Further more, the function $\sqrt{r_1/r_2} f_0$ belongs to $star(\mathcal{F}, 0)$ and satisfies that $T(\sqrt{r_1/r_2} f_0) \leq r_1$ Then

$$\begin{aligned}\psi(r_1) &= \mathbb{E}_\sigma R_n(\{f \in star(\mathcal{F}, 0) : T(f) \leq r_1\}) \geq \mathbb{E}_\sigma R_n(\sqrt{r_1/r_2} f_0) \\ &= \sqrt{r_1/r_2}\mathbb{E}_\sigma R_n(f_0) = \sqrt{r_1/r_2}\psi(r_2)\end{aligned}$$

$\square$

Estimating r^* from global information

Theorem 12

Let F be a class of {0,1}-valued functions with VC-dimension $d < \infty$. Then for all $K > 1$ and every $\delta > 0$, with probability at least $1 - \delta$, every $f \in \mathcal{F}$ satisfies

$$\mathbb{E}f \leq \hat{\mathbb{E}}f + cK\left(\frac{d \ln n/d}{n} + \frac{\ln 1/\delta}{n}\right).$$

Proof.

Define the sub-root function

$$\psi(r) = 10\mathbb{E}R_n(\{f \in star(\mathcal{F}, 0) : \mathbb{E}f^2 \leq 2\}) + \frac{11 \ln n}{n}.$$

If $r \geq \psi(r)$, Theorem 6 implies that, with probability at least $1 - 1/n$,

$$\{f \in star(\mathcal{F}, 0) : \mathbb{E}f^2 \leq r\} \subset \{f \in star(\mathcal{F}, 0) : \hat{\mathbb{E}}f^2 \leq 2r\},$$

and thus

Estimating r^* from global information

Proof.

$$\mathbb{E}R_n(\{f \in star(\mathcal{F},0) : \mathbb{E}f^2 \le r\}) \le \mathbb{E}R_n(\{f \in star(\mathcal{F},0) : \hat{\mathbb{E}}f^2 \le 2r\}) + 1/n$$

It follows that $r^* = \psi(r^*)$ satisfies

$$r^* = \psi(r^*) \le \mathbb{E}R_n(\{f \in star(\mathcal{F},0) : \hat{\mathbb{E}}f^2 \le 2r^*\}) + (1 + 11 \ln n)/n \quad (0.6)$$

Eqn (0.5) shows that

$$\mathbb{E}R_n(\{f \in star(\mathcal{F},0) : \hat{\mathbb{E}}f^2 \le 2r^*\}) \le \frac{C}{\sqrt{n}}\mathbb{E}\int_0^{\sqrt{2r^*}} \sqrt{\ln N(\epsilon, star(\mathcal{F},0), L_2(P_n))}d\epsilon$$

□

Estimating r^* from global information

Proof.

It is easy to see that we can construct an $\epsilon-$cover for star$(\mathcal{F}, 0)$ using an $\epsilon/2-$cover for $\mathcal{F}$ and an $\epsilon/2-$cover for the interval $[0, 1]$, which implies

$$\ln N(\epsilon, star(\mathcal{F}, 0), L_2(P_n)) \leq \ln N(\frac{\epsilon}{2}, \mathcal{F}, L_2(P_n))(\frac{2}{\epsilon} + 1)$$

Now, recall that for any distribution $\mathbb{P}$ and any class $\mathcal{F}$ with VC-dimension $d < \infty$,

$$\ln N(\frac{\epsilon}{2}, \mathcal{F}, L_2(P)) \leq cd \ln(\frac{1}{\epsilon}).$$

Therefore

$$\begin{aligned} \mathbb{E}R_n(\{f \in star(\mathcal{F}, 0) : \hat{\mathbb{E}}f^2 \leq 2r^*\}) &\leq \sqrt{\frac{cd}{n}} \int_0^{\sqrt{2r^*}} \sqrt{\ln(\frac{1}{\epsilon})} d\epsilon \\ &\leq \sqrt{cdr^* \ln(1/r^*)/n} \end{aligned}$$

Estimating r^* from global information

Proof.

Solve this equation, we have

$$r^* \leq \frac{cd \ln(n/d)}{n}.$$

$\square$