Rademacher Complexity and Concentration Inequalities

Mathematical Analysis of Machine Learning Algorithms
(Chapter 6)

Notations

Using the notations from Section 3.3, we are given a function class $\mathcal{G} = \{\phi(w, z) : w \in \Omega\}$, and are interested in the uniform convergence of training error

$$\phi(w, \mathcal{S}_n) = \frac{1}{n} \sum_{i=1}^{n} \phi(w, Z_i)$$

on a training data $\mathcal{S}_n = \{Z_1, \ldots, Z_n\} \sim \mathcal{D}^n$, to the test error

$$\phi(w, \mathcal{D}) = \mathbb{E}_{Z \sim \mathcal{D}} \phi(w, Z)$$

on the test data $\mathcal{D}$. In particular, in the general analysis of learning algorithms, we want to estimate the supremum of the associated empirical process:

$$\sup_{w \in \Omega} \left[\phi(w, \mathcal{D}) - \phi(w, \mathcal{S}_n) \right].$$

Uniform Convergence Complexity

We introduce the following definition of one-sided uniform convergence in expectation, which will be convenient in our analysis.

Definition 1 (Def 6.1)

Given an empirical process $\{\phi(w, \mathcal{S}_n) : w \in \Omega\}$, with $\mathcal{S}_n \sim \mathcal{D}^n$. Define the expected supremum of this empirical process as

$$\epsilon_n(\mathcal{G}, \mathcal{D}) = \mathbb{E}_{\mathcal{S}_n} \sup_{w \in \Omega} \left[\phi(w, \mathcal{D}) - \phi(w, \mathcal{S}_n)\right],$$

which will be referred to as the *uniform convergence complexity* of the function class $\mathcal{G}$.

Expected Oracle Inequality

Recall approximate ERM method

$$\phi(\hat{w}, \mathcal{S}_n) \leq \inf_{w \in \Omega} \phi(w, \mathcal{S}_n) + \epsilon'. \tag{1}$$

We have

Theorem 2 (Thm 6.2)

Consider $\phi(w, Z)$ *with* $Z \sim \mathcal{D}$. *Let* $\mathcal{S}_n \sim \mathcal{D}^n$ *be n iid samples from* $\mathcal{D}$. *Then the approximate ERM method of* (1) *satisfies*

$$\mathbb{E}_{\mathcal{S}_n} \phi(\hat{w}, \mathcal{D}) \leq \inf_{w \in \Omega} \phi(w, \mathcal{D}) + \epsilon' + \epsilon_n(\mathcal{G}, \mathcal{D}).$$

Proof of Theorem 2

Given any $w \in \Omega$, we have for each instance of training data $\mathcal{S}_n$

$$\begin{aligned}\phi(\hat{w}, \mathcal{D}) \leq & \phi(\hat{w}, \mathcal{S}_n) + \sup_{w \in \Omega} [\phi(w, \mathcal{D}) - \phi(w, \mathcal{S}_n)] \\ \leq & \phi(w, \mathcal{S}_n) + \epsilon' + \sup_{w \in \Omega} [\phi(w, \mathcal{D}) - \phi(w, \mathcal{S}_n)].\end{aligned}$$

Taking expectation with respect to $\mathcal{S}_n$, and note that w does not depend on $\mathcal{S}_n$, we obtain

$$\mathbb{E}_{\mathcal{S}_n} \phi(\hat{w}, \mathcal{D}) \leq \phi(w, \mathcal{D}) + \epsilon' + \mathbb{E}_{\mathcal{S}_n} \sup_{w \in \Omega} [\phi(w, \mathcal{D}) - \phi(w, \mathcal{S}_n)].$$

This implies the desired bound.

Rademacher Complexity

Definition 3 (One-sided Rademacher Complexity, Def 6.3)

Given $\mathcal{S}_n = \{Z_1, \ldots, Z_n\}$, the (one-sided) empirical Rademacher complexity of $\mathcal{G}$ is defined as

$$R(\mathcal{G}, \mathcal{S}_n) = \mathbb{E}_\sigma \sup_{w \in \Omega} \frac{1}{n} \sum_{i=1}^{n} \sigma_i \phi(w, Z_i),$$

where $\sigma_1, \ldots, \sigma_n$ are independent uniform $\{\pm 1\}$-valued Bernoulli random variables. Moreover, the expected Rademacher complexity is

$$R_n(\mathcal{G}, \mathcal{D}) = \mathbb{E}_{\mathcal{S}_n \sim \mathcal{D}^n} R(\mathcal{G}, \mathcal{S}_n).$$

Rademacher Complexity Bounds

Theorem 4 (Thm 6.4)

We have

$$\epsilon_n(\mathcal{G}, \mathcal{D}) \leq 2R_n(\mathcal{G}, \mathcal{D}).$$

Consequently, the approximate ERM method of (1) *satisfies*

$$\mathbb{E}_{\mathcal{S}_n}\phi(\hat{w}, \mathcal{D}) \leq \inf_{w \in \Omega} \phi(w, \mathcal{D}) + \epsilon' + 2R_n(\mathcal{G}, \mathcal{D}).$$

Proof of Theorem 4

Let $\mathcal{S}'_n = \{Z'_1, \ldots, Z'_n\} \sim \mathcal{D}^n$ be n iid samples from $\mathcal{D}$ that are independent of $\mathcal{S}_n$. We have

$$\begin{aligned}
\epsilon_n(\mathcal{G}, \mathcal{D}) =& \mathbb{E}_{\mathcal{S}_n \sim \mathcal{D}^n} \sup_{w \in \Omega} [\phi(w, \mathcal{D}) - \phi(w, \mathcal{S}_n)] \\
=& \mathbb{E}_{\mathcal{S}_n \sim \mathcal{D}^n} \sup_{w \in \Omega} [\mathbb{E}_{\mathcal{S}'_n \sim \mathcal{D}^n} \phi(w, \mathcal{S}'_n) - \phi(w, \mathcal{S}_n)] \\
\leq& \mathbb{E}_{(\mathcal{S}_n, \mathcal{S}'_n) \sim \mathcal{D}^{2n}} \sup_{w \in \Omega} [\phi(w, \mathcal{S}'_n) - \phi(w, \mathcal{S}_n)] \\
=& \mathbb{E}_{(\mathcal{S}_n, \mathcal{S}'_n) \sim \mathcal{D}^{2n}} \mathbb{E}_\sigma \sup_{w \in \Omega} \frac{1}{n} \sum_{i=1}^{n} [\sigma_i \phi(w, Z'_i) - \sigma_i \phi(w, Z_i)] \\
\leq& \mathbb{E}_{(\mathcal{S}_n, \mathcal{S}'_n) \sim \mathcal{D}^{2n}} [R(\mathcal{G}, \mathcal{S}_n) + R(\mathcal{G}, \mathcal{S}'_n)] = 2R_n(\mathcal{G}, \mathcal{D}).
\end{aligned}$$

This proves the desired bound.

Example

Example 5 (Expl 6.5)

Consider a (binary-valued) VC class $\mathcal{G}$ such that $\mathrm{VC}(\mathcal{G}) = d$. Consider $n \geq d$. Then Sauer's lemma implies that for any $\mathcal{S}_n$, the number of functions of $\phi \in \mathcal{G}$ on $\mathcal{S}_n$ is no more than $(en/d)^d$. We thus obtain (see Theorem 10)

$$R(\mathcal{G}, \mathcal{S}_n) \leq \sqrt{\frac{2d \ln(en/d)}{n}}.$$

This implies that the approximate ERM method of (1) satisfies

$$\mathbb{E}_{\mathcal{S}_n} \phi(\hat{w}, \mathcal{D}) \leq \inf_{w \in \Omega} \phi(w, \mathcal{D}) + \epsilon' + 2\sqrt{\frac{2d \ln(en/d)}{n}}.$$

Note: a better bound can be obtained using Theorem 5.6 and Theorem 6.25, which removes the $\ln n$ factor.

Example

Example 6 (Expl 6.12)

Consider regularized linear function class
$\mathcal{F}_{A,B} = \{\{f(w,x) = w^\top \psi(x) : \|w\|_2 \le A, \|\psi(x)\|_2 \le B\}$. $\forall \lambda > 0$:

$$M(\lambda) = \mathbb{E}_\sigma \sup_{w \in \mathbb{R}^d} \left[\frac{1}{n} \sum_{i=1}^n \sigma_i w^\top \psi(X_i) - \frac{\lambda}{4} \|w\|_2^2 \right]$$
$$= \frac{1}{\lambda} \mathbb{E}_\sigma \left\| \frac{1}{n} \sum_{i=1}^n \sigma_i \psi(X_i) \right\|_2^2 = \frac{1}{\lambda n^2} \sum_{i=1}^n \|\psi(X_i)\|_2^2.$$

Let $\mathcal{F}_{A,B} = \{f(w,x) = w^\top \psi(x) : \|w\|_2 \le A, \|\psi(x)\|_2 \le B\}$, then

$$R(\mathcal{F}_{A,B}, \mathcal{S}_n) \le \inf_{\lambda > 0} \left[M(\lambda) + \frac{\lambda}{4} A^2 \right] \le \inf_{\lambda > 0} \left[\frac{B^2}{\lambda n} + \frac{\lambda}{4} A^2 \right] = AB/\sqrt{n}.$$

Concentration Inequality

Theorem 7 (McDiarmid's Inequality, Thm 6.16)

Consider n independent random variables $X_1, \ldots, X_n$, and a real-valued function $f(X_1, \ldots, X_n)$ that satisfies the following inequality

$$\sup_{x_1,\ldots,x_n,x_i'} |f(x_1, \ldots, x_n) - f(x_1, \ldots, x_{i-1}, x_i', x_{i+1}, \ldots, x_n)| \leq c_i$$

for all $1 \leq i \leq n$. Then for all $\epsilon > 0$:

$$\Pr\left[f(X_1, \ldots, X_n) \geq \mathbb{E}f(X_1, \ldots, X_n) + \epsilon\right] \leq \exp\left(\frac{-2\epsilon^2}{\sum_{i=1}^n c_i^2}\right).$$

Similarly:

$$\Pr\left[f(X_1, \ldots, X_n) \leq \mathbb{E}f(X_1, \ldots, X_n) - \epsilon\right] \leq \exp\left(\frac{-2\epsilon^2}{\sum_{i=1}^n c_i^2}\right).$$

Proof of Theorem 7 (I/III)

Let $X_k^l = \{X_k, \ldots, X_l\}$. Consider X_1^n, and for some $1 \leq k \leq n$, we use the simplified notation $\tilde{X}_1^n = \{X_1, \ldots, X_{k-1}, \tilde{X}_k, X_{k+1}, X_n\}$. Then

$$|\mathbb{E}_{X_{k+1}^n} f(X_1^n) - \mathbb{E}_{X_{k+1}^n} f(\tilde{X}_1^n)| \leq c_k.$$

We now consider $\mathbb{E}_{X_{k+1}^n} f(X_1^n)$ as a random variable depending on X_k, conditioned on X_1^{k-1}. We have:

$$\ln \mathbb{E}_{X_k} \exp[\lambda \mathbb{E}_{X_{k+1}^n} f(X_1^n)] \leq \lambda \mathbb{E}_{X_k^n} f(X_1^n) + \lambda^2 c_k^2/8.$$

Proof of Theorem 7 (II/III)

Now we may exponentiate the above inequality, and take expectation with respect to X_1^{k-1} to obtain

$$\mathbb{E}_{X_1^k} \exp[\lambda \mathbb{E}_{X_{k+1}^n} f(X_1^n)] \leq \mathbb{E}_{X_1^{k-1}} \exp[\lambda \mathbb{E}_{X_k^n} f(X_1^n) + \lambda^2 c_k^2/8].$$

By taking logarithm, we obtain

$$\ln \mathbb{E}_{X_1^k} \exp[\lambda \mathbb{E}_{X_{k+1}^n} f(X_1^n)] \leq \ln \mathbb{E}_{X_1^{k-1}} \exp[\lambda \mathbb{E}_{X_k^n} f(X_1^n)] + \lambda^2 c_k^2/8.$$

By summing from $k = 1$ to n, and canceling redundant terms:

$$\ln \mathbb{E}_{X_1^n} \exp[\lambda f(X_1^n)] \leq \lambda \mathbb{E}_{X_1^n} f(X_1^n) + \lambda^2 \sum_{k=1}^{n} c_k^2/8. \tag{2}$$

Proof of Theorem 7 (III/III)

Let

$$\delta = \Pr\left[f(X_1^n) \geq \mathbb{E}_{X_1^n} f(X_1^n) + \epsilon\right].$$

Using Markov's inequality, we have for all positive λ

$$\delta \leq e^{-\lambda(\mathbb{E}_{X_1^n} f(X_1^n)+\epsilon)} \mathbb{E}_{X_1^n} e^{\lambda f(X_1^n)} \leq \exp\left[-\lambda\epsilon + \frac{\lambda^2}{8}\sum_{k=1}^{n} c_k^2\right].$$

Since $\lambda > 0$ is arbitrary, we conclude that

$$\ln\delta \leq \inf_{\lambda \geq 0}\left[\frac{\lambda^2}{8}\sum_{k=1}^{n} c_k^2 - \lambda\epsilon\right] = -\frac{2\epsilon^2}{\sum_{k=1}^{n} c_k^2}.$$

This implies the theorem.

Example of McDiarmid's Inequality

McDiarmid's inequality is referred to as *concentration inequality* because it states that the sample dependent quantity $f(X_1, \ldots, X_n)$ does not deviate significantly from its expectation $\mathbb{E}f(X_1, \ldots, X_n)$.

Additive Chernoff Bound

Note that if we take

$$f(x_1, \ldots, x_n) = \frac{1}{n} \sum_{i=1}^{n} x_i,$$

and assume that $x_i \in [0, 1]$, then we can take $c_i = 1/n$ in McDiarmid's inequality. This implies

$$\Pr\left[f(X_1, \ldots, X_n) \geq \mathbb{E}f(X_1, \ldots, X_n) + \epsilon\right] \leq \exp\left(-2n\epsilon^2\right).$$

McDiarmid's inequality is a generalization of additive Chernoff bound.

Uniform Convergence

We can apply McDiarmid's inequality to obtain the following uniform convergence result in large probability.

Corollary 8 (Simplification with $h(\cdot) = 0$, Cor 6.19)

Assume that for some $M \geq 0$:

$$\sup_{w \in \Omega} \sup_{z, z'} \left[\phi(w, z) - \phi(w, z')\right] \leq M.$$

Then with probability at least $1 - \delta$: for all $w \in \Omega$,

$$\begin{aligned} \phi(w, \mathcal{D}) \leq & \phi(w, \mathcal{S}_n) + \epsilon_n(\mathcal{G}, \mathcal{D}) + M\sqrt{\frac{\ln(1/\delta)}{2n}} \\ \leq & \phi(w, \mathcal{S}_n) + 2R_n(\mathcal{G}, \mathcal{D}) + M\sqrt{\frac{\ln(1/\delta)}{2n}}. \end{aligned}$$

Proof of Theorem 8

Consider $\mathcal{S}_n = \{Z_1, \ldots, Z_n\}$ and $\mathcal{S}'_n = \{Z_1, \ldots, Z_{i-1}, Z'_i, Z_{i+1}, \ldots, Z_n\}$. Let $f(\mathcal{S}_n) = \sup_{w \in \Omega}[\phi(w, \mathcal{D}) - \phi(w, \mathcal{S}_n)]$. For simplicity, we assume that the sup can be achieved at $\hat{w}$ as

$$\hat{w} = \arg\max_{w \in \Omega}[\phi(w, \mathcal{D}) - \phi(w, \mathcal{S}_n)].$$

Then

$$\begin{aligned}
& f(\mathcal{S}_n) - f(\mathcal{S}'_n) \\
=& [\phi(\hat{w}, \mathcal{D}) - \phi(\hat{w}, \mathcal{S}_n)] - \sup_{w \in \Omega}[\phi(w, \mathcal{D}) - \phi(w, \mathcal{S}'_n)] \\
\leq& [\phi(\hat{w}, \mathcal{D}) - \phi(\hat{w}, \mathcal{S}_n)] - [\phi(\hat{w}, \mathcal{D}) - \phi(\hat{w}, \mathcal{S}'_n)] \leq M/n.
\end{aligned}$$

Similarly, $f(\mathcal{S}'_n) - f(\mathcal{S}_n) \leq M/n$. Therefore we may take $c_i = M/n$ in Theorem 7, which implies the first desired result. The second bound follows from the estimate $\epsilon_n(\mathcal{G}, \mathcal{D}) \leq 2R_n(\mathcal{G}, \mathcal{D})$ of Theorem 4.

Oracle Inequality

Corollary 9 (Simplification with $h(\cdot) = 0$, Cor 6.21)

Assume that for some $M \geq 0$*:*

$$\sup_{w \in \Omega} \sup_{z, z'} \left[\phi(w, z) - \phi(w, z') \right] \leq M.$$

Then the approximate ERM method

$$\phi(\hat{w}, \mathcal{S}_n) \leq \min_{w \in \Omega} \phi(w, \mathcal{S}_n) + \epsilon' \tag{3}$$

satisfies the following oracle inequality. With probability at least $1 - \delta$*:*

$$\begin{aligned} \phi(\hat{w}, \mathcal{D}) &\leq \inf_{w \in \Omega} \phi(w, \mathcal{D}) + \epsilon' + \epsilon_n(\mathcal{G}, \mathcal{D}) + 2M \sqrt{\frac{\ln(2/\delta)}{2n}} \\ &\leq \inf_{w \in \Omega} \phi(w, \mathcal{D}) + \epsilon' + 2R_n(\mathcal{G}, \mathcal{D}) + 2M \sqrt{\frac{\ln(2/\delta)}{2n}}. \end{aligned}$$

Proof of Corollary 9

Given any $w \in \Omega$, from the Chernoff bound, we know that with probability $1 - \delta/2$,

$$\phi(w, \mathcal{S}_n) \leq \phi(w, \mathcal{D}) + M\sqrt{\frac{\ln(2/\delta)}{2n}}. \tag{4}$$

Taking the union bound with the inequality of Corollary 8 at $\delta/2$, we obtain at probability $1 - \delta$,

$$\begin{aligned} \phi(\hat{w}, \mathcal{D}) \leq & \phi(\hat{w}, \mathcal{S}_n) + \epsilon_n(\mathcal{G}, \mathcal{D}) + M\sqrt{\frac{\ln(2/\delta)}{2n}} \\ \leq & \phi(w, \mathcal{D}) + \epsilon' + \epsilon_n(\mathcal{G}, \mathcal{D}) + 2M\sqrt{\frac{\ln(2/\delta)}{2n}}. \end{aligned}$$

In the above derivation, the first inequality used Corollary 8. The second inequality used (4). This proves the first desired bound. The second desired bound employs Theorem 4.

Estimating Rademacher Complexity

Theorem 10 (First Inequality of Thm 6.23)

If $\mathcal{G}$ is a finite function class with $|\mathcal{G}| = N$, then

$$R(\mathcal{G}, \mathcal{S}_n) \leq \sup_{g \in \mathcal{G}} \|g\|_{L_2(\mathcal{S}_n)} \cdot \sqrt{\frac{2 \ln N}{n}}.$$

Proof of Theorem 10

Let $B = \sup_{g\in\mathcal{G}} \|g\|_{L_2(\mathcal{S}_n)}$. Then we have for all $\lambda > 0$:

$$
\begin{aligned}
R(\mathcal{G}, \mathcal{S}_n) =& \mathbb{E}_\sigma \sup_{g\in\mathcal{G}} \frac{1}{n} \sum_{i=1}^n \sigma_i g(Z_i) \\
\overset{(a)}{\leq} & \mathbb{E}_\sigma \frac{1}{\lambda n} \ln \sum_{g\in\mathcal{G}} \exp\left[\lambda \sum_{i=1}^n \sigma_i g(Z_i)\right] \\
\overset{(b)}{\leq} & \frac{1}{\lambda n} \ln \mathbb{E}_\sigma \sum_{g\in\mathcal{G}} \exp\left[\lambda \sum_{i=1}^n \sigma_i g(Z_i)\right] \\
=& \frac{1}{\lambda n} \ln \sum_{g\in\mathcal{G}} \prod_{i=1}^n \mathbb{E}_{\sigma_i} \exp\left[\lambda \sigma_i g(Z_i)\right] \\
\overset{(c)}{\leq} & \frac{1}{\lambda n} \ln \sum_{g\in\mathcal{G}} \prod_{i=1}^n \exp[\lambda^2 g(Z_i)^2/2] \leq \frac{1}{\lambda n} \ln N \exp[\lambda^2 n B^2/2].
\end{aligned}
$$

Now we can obtain the desired bound by optimizing over $\lambda > 0$.

Compare with Covering Number Results

Consider $\phi(w, Z) \in [0, 1]$ and $|\mathcal{G}| = N$. With probability $1 - \delta$. We have the following uniform convergence results for all w. If we use the union of Chernoff bound (covering number) method, then

$$\phi(w, \mathcal{D}) \leq \phi(w, \mathcal{S}_n) + \sqrt{\frac{\ln(N/\delta)}{2n}},$$

which implies that

$$\phi(w, \mathcal{D}) \leq \phi(w, \mathcal{S}_n) + \sqrt{\frac{\ln(N)}{2n}} + \sqrt{\frac{\ln(1/\delta)}{2n}}.$$

Rademacher bound from Corollary 8 (with Rademacher complexity estimate from Theorem 10):

$$\phi(w, \mathcal{D}) \leq \phi(w, \mathcal{S}_n) + 4\sqrt{\frac{\ln(N)}{2n}} + \sqrt{\frac{\ln(1/\delta)}{2n}},$$

which leads to similar result.

Chaining

Chapter 4 employs empirical L_1 covering number bound to obtain uniform convergence of

$$\inf_{\epsilon>0}\left[\epsilon+\sup_{\mathcal{S}_n}\sqrt{\frac{\ln N(\epsilon/2, G, L_1(\mathcal{S}_n))}{n}}\right].$$

This can be improved by considering multiple approximation scales with empirical L_2 covering numbers, instead of a single scale.

Theorem 11 (Thm 6.25)

We have

$$R(\mathcal{G},\mathcal{S}_n)\le\inf_{\epsilon\ge 0}\left[4\epsilon+12\int_{\epsilon}^{\infty}\sqrt{\frac{\ln N(\epsilon',\mathcal{G},L_2(\mathcal{S}_n))}{n}}d\epsilon'\right].$$

Proof of Theorem 11 (I/II)

Let $B = \sup_{g\in\mathcal{G}} \|g\|_{L_2(\mathcal{S}_n)}$, and let $\epsilon_\ell = 2^{-\ell}B$ for $\ell = 0, 1, \ldots$. Let $\mathcal{G}_\ell$ be an ϵ_ℓ-cover of $\mathcal{G}$ with metric $L_2(\mathcal{S}_n)$, and $N_\ell = |\mathcal{G}_\ell| = N(\epsilon_\ell, \mathcal{G}, L_2(\mathcal{S}_n))$.
We may let $\mathcal{G}_0 = \{0\}$ at scale $\epsilon_0 = B$.
For each $g \in \mathcal{G}$, we consider $g_\ell(g) \in \mathcal{G}_\ell$ so that $\|g - g_\ell(g)\|_{L_2(\mathcal{S}_n)} \le \epsilon_\ell$.
The key idea in chaining is to rewrite $g \in \mathcal{G}$ using the following multi-scale decomposition:

$$g = (g - g_L(g)) + \sum_{\ell=1}^{L}(g_\ell(g) - g_{\ell-1}(g)).$$

We also have

$$\|g_\ell(g) - g_{\ell-1}(g)\|_{L_2(\mathcal{S}_n)} \le \|g_\ell(g) - g\|_{L_2(\mathcal{S}_n)} + \|g_{\ell-1}(g) - g\|_{L_2(\mathcal{S}_n)} \le 3\epsilon_\ell. \tag{5}$$

The number of distinct $g_\ell(g) - g_{\ell-1}(g)$ is no more than $N_\ell N_{\ell-1}$.

Proof of Theorem 11 (II/II)

It implies that

$$R(\mathcal{G},\mathcal{S}_n) = \mathbb{E}_\sigma \sup_{g\in\mathcal{G}} \frac{1}{n}\sum_{i=1}^n \sigma_i \left[(g - g_L(g))(Z_i) + \sum_{\ell=1}^L (g_\ell(g) - g_{\ell-1}(g))(Z_i)\right]$$

$$\le \mathbb{E}_\sigma \sup_{g\in\mathcal{G}} \frac{1}{n}\sum_{i=1}^n \sigma_i (g - g_L(g))(Z_i) + \sum_{\ell=1}^L \mathbb{E}_\sigma \sup_{g\in\mathcal{G}} \frac{1}{n}\sum_{i=1}^n \sigma_i (g_\ell(g) - g_{\ell-1}(g))(Z_i)$$

$$\overset{(a)}{\le} \epsilon_L + \sum_{\ell=1}^L \sup_{g\in\mathcal{G}} \|g_\ell(g) - g_{\ell-1}(g)\|_{L_2(\mathcal{S}_n)} \sqrt{\frac{2\ln[N_\ell N_{\ell-1}]}{n}}$$

$$\overset{(b)}{\le} \epsilon_L + 3\sum_{\ell=1}^L \epsilon_\ell \sqrt{\frac{2\ln[N_\ell N_{\ell-1}]}{n}}$$

$$\le \epsilon_L + 12\sum_{\ell=1}^L (\epsilon_\ell - \epsilon_{\ell+1}) \sqrt{\frac{\ln[N_\ell]}{n}}$$

$$\le \epsilon_L + 12\int_{\epsilon_L/2}^\infty \sqrt{\frac{\ln N(\epsilon',\mathcal{G},L_2(\mathcal{S}_n))}{n}}\, d\epsilon'.$$

VC-Class Example: Rademacher Complexity

Example 12

If a binary-valued function class $\mathcal{G}$ (or a VC-subgraph class with values in $[0, 1]$) has VC-dimension d, then (see Corollary 5.7)

$$\ln N_2(\epsilon, \mathcal{G}, n) \leq 1 + \ln(d + 1) + d \ln(2e/\epsilon^2).$$

Therefore

$$\begin{aligned} 12 \int_0^\infty \sqrt{\ln N_2(\epsilon, \mathcal{G}, n)} d\epsilon \leq & 12 \int_0^{0.5} \sqrt{1 + \ln(d + 1) + d \ln(2e/\epsilon^2)} d\epsilon \\ \leq & 16\sqrt{d}. \end{aligned}$$

It follows that

$$R(\mathcal{G}, \mathcal{S}_n) \leq \frac{16\sqrt{d}}{\sqrt{n}}.$$

VC-Class Example: Uniform Convergence

The Rademacher complexity result of VC-subgraph class and Corollary 8 imply the following uniform convergence result.

Uniform Convergence of VC-subgraph Class

Let $\mathcal{G} = \{\phi(w, \cdot) : w \in \Omega\}$ be a VC-subgraph class with VC-dimension d. With probability at least $1 - \delta$, for all $w \in \Omega$,

$$\phi(w, \mathcal{D}) \leq \phi(w, \mathcal{S}_n) + \frac{32\sqrt{d}}{\sqrt{n}} + \sqrt{\frac{\ln(1/\delta)}{2n}}.$$

This bound removes a $\ln n$ factor from the additive uniform convergence bound in Theorem 4.17 which employs the L_1 empirical covering number analysis.

Example: Nonparameteric Function Class

Example 13

If $\ln N_2(\epsilon, \mathcal{G}, n) \leq 1/\epsilon^q$ for $q \in (0, 2)$, then

$$\int_0^\infty \sqrt{\ln N_2(\epsilon, \mathcal{G}, n)} d\epsilon < \infty.$$

Therefore there exists $C > 0$ such that

$$R(\mathcal{G}, \mathcal{S}_n) \leq \frac{C}{\sqrt{n}}.$$

If $\ln N_2(\epsilon, \mathcal{G}, n) \leq 1/\epsilon^q$ for $q > 2$, then

$$R(\mathcal{G}, \mathcal{S}_n) \leq O\left(\inf_{\epsilon > 0}\left(\epsilon + \frac{\epsilon^{1-q/2}}{\sqrt{n}}\right)\right) = O(n^{-1/q}).$$

This implies a convergence slower than $1/\sqrt{n}$.

Lipschitz Composition

Let $\{\phi_i\}$ be a set of functions, each characterized by a Lipschitz constant γ_i, namely

$$|\phi_i(\theta) - \phi_i(\theta')| \leq \gamma_i |\theta - \theta'|.$$

Then the result implies a bound on the Rademacher complexity of the function composition $\phi \circ f$.

Theorem 14 (Simplified with $h(\cdot) = 0$, Thm 6.28)

Let $\{\phi_i\}_{i=1}^n$ be functions with Lipschitz constants $\{\gamma_i\}_{i=1}^n$ respectively. That is, $\forall i \in [n]$:

$$|\phi_i(\theta) - \phi_i(\theta')| \leq \gamma_i |\theta - \theta'|.$$

Then for any $\mathcal{S}_n = \{Z_1, \ldots, Z_n\} \subset \mathcal{Z}^n$, we have

$$\mathbb{E}_\sigma \sup_{f \in \mathcal{F}} \left[\sum_{i=1}^n \sigma_i \phi_i(f(Z_i)) \right] \leq \mathbb{E}_\sigma \sup_{f \in \mathcal{F}} \left[\sum_{i=1}^n \sigma_i \gamma_i f(Z_i) \right].$$

Proof of Theorem 14

The result is a direct consequence of the Lemma 15, where we simply set $c(w) = 0$, $g_i(w) = \phi_i(f(Z_i))$, and $\tilde{g}_i(w) = \gamma_i f(Z_i)$.

Lemma 15 (Rademacher comparison lemma, Lem 6.29)

Let $\{g_i(w)\}$ and $\{\tilde{g}_i(w)\}$ be sets of functions defined for all w in some domain Ω. If for all i, w, w',

$$|g_i(w) - g_i(w')| \leq |\tilde{g}_i(w) - \tilde{g}_i(w')|,$$

then for any function $c(w)$,

$$\mathbb{E}_\sigma \sup_{w \in \Omega} \left[c(w) + \sum_{i=1}^{n} \sigma_i g_i(w) \right] \leq \mathbb{E}_\sigma \sup_{w \in \Omega} \left[c(w) + \sum_{i=1}^{n} \sigma_i \tilde{g}_i(w) \right].$$

Proof of Lemma 15 (I/II)

We prove this result by induction. The result holds for $n = 0$. Assume that the result holds for $n = k$, then when $n = k + 1$, we have:

$$
\begin{aligned}
&\mathbb{E}_{\sigma_1,\dots,\sigma_{k+1}} \sup_{w} \left[c(w) + \sum_{i=1}^{k+1} \sigma_i g_i(w) \right] \\
&= \mathbb{E}_{\sigma_1,\dots,\sigma_k} \sup_{w_1,w_2} \left[\frac{c(w_1) + c(w_2)}{2} + \sum_{i=1}^{k} \sigma_i \frac{g_i(w_1) + g_i(w_2)}{2} \right. \\
&\qquad\qquad \left. + \frac{g_{k+1}(w_1) - g_{k+1}(w_2)}{2} \right] \\
&= \mathbb{E}_{\sigma_1,\dots,\sigma_k} \sup_{w_1,w_2} \left[\frac{c(w_1) + c(w_2)}{2} + \sum_{i=1}^{k} \sigma_i \frac{g_i(w_1) + g_i(w_2)}{2} \right. \\
&\qquad\qquad \left. + \frac{|g_{k+1}(w_1) - g_{k+1}(w_2)|}{2} \right] \\
&= A.
\end{aligned}
$$

Proof of Lemma 15 (II/II)

We continue from the previous derivation, with:

$$
\begin{aligned}
A &\leq \mathbb{E}_{\sigma_1,\ldots,\sigma_k} \sup_{w_1,w_2} \left[\frac{c(w_1)+c(w_2)}{2} + \sum_{i=1}^{k} \sigma_i \frac{g_i(w_1)+g_i(w_2)}{2} \right. \\
&\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad \left. + \frac{|\tilde{g}_{k+1}(w_1) - \tilde{g}_{k+1}(w_2)|}{2} \right] \\
&= \mathbb{E}_{\sigma_1,\ldots,\sigma_k} \sup_{w_1,w_2} \left[\frac{c(w_1)+c(w_2)}{2} + \sum_{i=1}^{k} \sigma_i \frac{g_i(w_1)+g_i(w_2)}{2} \right. \\
&\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad \left. + \frac{\tilde{g}_{k+1}(w_1) - \tilde{g}_{k+1}(w_2)}{2} \right] \\
&= \mathbb{E}_{\sigma_1,\ldots,\sigma_k} \mathbb{E}_{\sigma_{k+1}} \sup_{w} \left[c(w) + \sigma_{k+1} \tilde{g}_{k+1}(w) + \sum_{i=1}^{k} \sigma_i g_i(w) \right] \\
&\leq \mathbb{E}_{\sigma_1,\ldots,\sigma_k} \mathbb{E}_{\sigma_{k+1}} \sup_{w} \left[c(w) + \sigma_{k+1} \tilde{g}_{k+1}(w) + \sum_{i=1}^{k} \sigma_i \tilde{g}_i(w) \right].
\end{aligned}
$$

The last inequality follows from the induction hypothesis.

Example

Example 16 (Variation of Expl 6.30)

Consider the regularized linear prediction functions

$$\mathcal{F}_{A,B} = \{\{f(w,x) = w^\top \psi(x) : \|w\|_2 \le A, \|\psi(x)\|_2 \le B\}$$

in Example 6. Consider smoothed classification loss function

$$L(f(x), y) = \min(1, \max(0, 1 - \gamma f(x) y))$$

for some $\gamma > 0$.
Let

$$\mathcal{G} = \{L(f(w,x), y) : f \in \mathcal{F}_{A,B}\}.$$

Then $L(f, y)$ is γ Lipschitz in f. We obtain from Theorem 14:

$$R(\mathcal{G}, \mathcal{S}_n) \le \gamma R(\mathcal{F}_{A,B}, \mathcal{S}_n) \le \gamma AB/\sqrt{n}.$$

Lipschitz Loss: Uniform Convergence

Theorem 17 (First Inequality of Thm 6.31)

Consider real-valued function class $\mathcal{F} = \{f(w, \cdot) : w \in \Omega\}$*, and*

$$\mathcal{G} = \{\phi(w, z) = L(f(w, x), y) : w \in \Omega, z = (x, y)\}.$$

Assume that $L(f, y)$ *is* γ*-Lipschitz in* f*:* $|L(f, y) - L(f', y)| \leq \gamma |f - f'|$*, and*

$$\sup_{(x,y),(x',y')} |L(f(w, x), y) - L(f(w, x'), y')| \leq M.$$

Let $\mathcal{S}_n \sim \mathcal{D}^n$*. With probability at least* $1 - \delta$*, for all* $w \in \Omega$*:*

$$\mathbb{E}_{\mathcal{D}} L(f(w, X), Y) \leq \frac{1}{n} \sum_{i=1}^{n} L(f(w, X_i), Y_i) + 2\gamma R_n(\mathcal{F}, \mathcal{D}) + M \sqrt{\frac{\ln(1/\delta)}{2n}}.$$

Lipschitz Loss: Oracle Inequality

Theorem 18 (Simplified Second Inequality of Thm 6.31)

Under the conditions of Theorem 17, and consider the approximate regularized ERM method (3) *with*

$$\phi(w, z) = L(f(w, x), y).$$

We have with probability at least $1 - \delta$*:*

$$\mathbb{E}_{\mathcal{D}} L(f(\hat{w}, X), Y) \leq \inf_{w \in \Omega} \mathbb{E}_{\mathcal{D}} L(f(w, X), Y) + \epsilon' + 2\gamma R_n(\mathcal{F}, \mathcal{D}) + M\sqrt{\frac{2\ln(2/\delta)}{n}},$$

where $\mathcal{F} = \{f(w, \cdot) : w \in \Omega\}$*.*

Talagrand's Concentration Inequality

Talagrand's concentration inequality is similar to Bernstein inequality, which is needed to derive faster than $1/\sqrt{n}$ concentration rate.

Corollary 19 (Cor 6.34)

Consider a real valued function class $\mathcal{F} = \{f(z) : \mathcal{Z} \to \mathbb{R}\}$. Let $\mathcal{D}$ be a distribution on $\mathcal{Z}$. Assume that there exists $M, \sigma > 0$ so that $\forall f \in \mathcal{F}$, $\sigma^2 \geq \mathrm{Var}_{Z\sim\mathcal{D}}[f(Z)]$, and $\sup_{z' \in \mathcal{Z}}[\mathbb{E}_{Z\sim\mathcal{D}} f(Z) - f(z')] \leq M$. Let $\mathcal{S}_n = \{Z_1, \ldots, Z_n\}$ be n independent random variables from $\mathcal{D}$. Then with probability at least $1-\delta$ over $\mathcal{S}_n$, for all $f \in \mathcal{F}$,

$$
\begin{aligned}
&\mathbb{E}_{Z\sim\mathcal{D}} f(Z) - \frac{1}{n}\sum_{i=1}^{n} f(Z_i) \\
\leq &\epsilon_n(\mathcal{F}, \mathcal{D}) + \sqrt{\frac{(4M\epsilon_n(\mathcal{F}, \mathcal{D}) + 2\sigma^2)\ln(1/\delta)}{n}} + \frac{M\ln(1/\delta)}{3n} \\
\leq &2\epsilon_n(\mathcal{F}, \mathcal{D}) + \sqrt{\frac{2\sigma^2 \ln(1/\delta)}{n}} + \frac{4M\ln(1/\delta)}{3n},
\end{aligned}
$$

where $\epsilon_n(\mathcal{F}, \mathcal{D})$ is Definition 1.

Fast Rate for Least Squares Regression (Expl 6.49)

Consider a function class $\mathcal{F}$ and the ERM method for least squares regression:

$$\hat{f} = \arg\min_{f \in \mathcal{F}} \frac{1}{n} \sum_{i=1}^{n} (f(X_i) - Y_i)^2,$$

where $Z_i = (X_i, Y_i)$ are iid samples from $\mathcal{D}$. Assume that $|f(X) - Y| \in [0, 1]$ for all X and Y. Example 3.18 implies that the loss function $\phi(f, Z) = [(f(X) - Y)^2 - (f_*(X) - Y)^2]$ satisfies the variance condition if the true regression function $f_* \in \mathcal{F}$. Assume also that the empirical covering number of $\mathcal{F}$ satisfies:

$$\ln N_2(\epsilon, \mathcal{F}, n) \leq \frac{c}{\epsilon^p} \quad (6)$$

for some constant $c > 0$ and $p > 0$.

Fast Rate for Least Squares Regression (cont)

We consider the following two situations: $p \in (0, 2)$ and $p \geq 2$.

- $p \in (0, 2)$. We obtain with probability at least $1 - \delta$:

$$\mathbb{E}_{\mathcal{D}} L(\hat{f}(X), Y) \leq \mathbb{E}_{\mathcal{D}} L(f_*(X), Y) + O\left(n^{-2/(2+p)} + \frac{\ln(1/\delta)}{n}\right).$$

- $p > 2$. The entropy integral of Theorem 11 implies that

$$R_n(\mathcal{F}^h(b), \mathcal{D}) \leq \frac{\tilde{c}_1}{n^{1/p}}$$

for some constant $\tilde{c}_1$. We thus obtain a rate of convergence of

$$\bar{r}^h(\alpha, \mathcal{F}, \mathcal{D}) = O\left(n^{-1/p}\right)$$

for local Rademacher complexity of Section 6.5. It can be shown that this is the same rate as what we can obtain from the standard non-localized Rademacher complexity.

Summary (Chapter 6)

- Uniform Convergence Complexity
- Expected Uniform Convergence and Expected Oracle Inequality
- Rademacher Complexity
- Concentration Inequality
- High Probability Uniform Convergence and Oracle Inequality
- Estimate Rademacher complexity
- Chaining and Dudley's entropy Integral estimate
- Composition with Lipschitz function and comparison lemma