Lower Bounds and Minimax Analysis

Mathematical Analysis of Machine Learning Algorithms
(Chapter 12)

Lower Bounds for Empirical Process

Consider empirical processes associated with a function family $\mathcal{F} = \{f(w, z) : w \in \Omega\}$, defined on the empirical measure $\mathcal{S}_n = \{Z_1, \ldots, Z_n\}$.

Definition 1 (Gaussian Complexity)

The empirical Gaussian complexity of $\mathcal{F}$ is defined as

$$G(\mathcal{F}, \mathcal{S}_n) = \mathbb{E}_g \sup_{f \in \mathcal{F}} \frac{1}{n} \sum_{i=1}^{n} g_i f(Z_i),$$

where $[g_1, \ldots, g_n]$ are independent standard normal random variables: $g_i \sim N(0, 1)$ for $i = 1, \ldots, n$.

Gaussian versus Rademacher Complexity

The following result shows that Gaussian complexity and Rademacher complexity are equivalent up to a logarithmic factor in n.

Proposition 2 ([Bartlett et al., 2002], Prop 12.2)

There exists an absolute constant $C > 0$ such that if $\mathcal{F} = -\mathcal{F}$, then

$$C^{-1} R(\mathcal{F}, \mathcal{S}_n) \leq G(\mathcal{F}, \mathcal{S}_n) \leq C \ln n \, R(\mathcal{F}, \mathcal{S}_n).$$

Upper bounds for both Rademacher and Gaussian complexities can be obtained from covering numbers and Dudley's entropy integral (chaining).

Lower bound for Gaussian complexity, called Sudakov Minoration, can be obtained via covering numbers.

Slepian's Lemma

Lemma 3 (Slepian's Lemma)

Let $[X_1, \ldots, X_n]$ *and* $[Y_1, \ldots, Y_n]$ *denote two zero-mean multivariate normal random vectors. Assume that*

$$\forall i \neq j, \qquad \mathbb{E}(X_i - X_j)^2 \geq \mathbb{E}(Y_i - Y_j)^2.$$

Then

$$\mathbb{E} \max_i X_i \geq \mathbb{E} \max_i Y_i.$$

Sudakov Minoration

Theorem 4 (Sudakov Minoration, Thm 12.4)

For any $\epsilon > 0$*:*

$$\sqrt{\ln M(\epsilon, \mathcal{F}, L_2(\mathcal{S}_n))} \leq \frac{2\sqrt{n} G(\mathcal{F}, \mathcal{S}_n)}{\epsilon} + 1.$$

Proof of Theorem 4 (I/III)

Let $\mathcal{F}_M = \{f_1, \ldots, f_M\} \subset \mathcal{F}$ be an ϵ packing subset of $\mathcal{F}$ under the $L_2(\mathcal{S}_n)$ metric. Consider independent standard Gaussian random variables $[g_1, \ldots, g_n]$.

$$\mathbb{E} \sup_{f \in \mathcal{F}} \frac{1}{n} \sum_{i=1}^{n} g_i f(Z_i) \geq \mathbb{E} \sup_{j \in [M]} \frac{1}{n} \sum_{i=1}^{n} g_i f_j(Z_i).$$

Let $g'_1, \ldots, g'_M$ be independent zero-mean normal random variables with variance $\epsilon^2/(2n)$ each. We then have for each $j \neq k$:

$$\begin{aligned}
&\mathbb{E} \left(\frac{1}{n} \sum_{i=1}^{n} g_i f_j(Z_i) - \frac{1}{n} \sum_{i=1}^{n} g_i f_k(Z_i) \right)^2 \\
=&\frac{1}{n^2} \sum_{i=1}^{n} (f_j(Z_i) - f_k(Z_i))^2 \\
\geq&\frac{1}{n} \epsilon^2 = \mathbb{E} \, (g'_j - g'_k)^2.
\end{aligned}$$

Proof of Theorem 4 (II/III)

Using Slepian's lemma, we have

$$G(\mathcal{F}, L_2(\mathcal{S}_n)) \geq \mathbb{E} \sup_j g'_j.$$

We know from Theorem 2.1 that for all j and $z \geq 0$,

$$\Pr\left(g'_j \leq \frac{\epsilon z}{\sqrt{2n}}\right) \leq 1 - 0.5e^{-(z+1)^2/2}$$
$$\Pr\left(g'_j \leq -\frac{\epsilon z}{\sqrt{2n}}\right) \leq 0.5e^{-z^2/2},$$

and thus

$$\Pr\left(\sup_j g'_j \leq \frac{\epsilon z}{\sqrt{2n}}\right) \leq (1 - 0.5e^{-(z+1)^2/2})^M$$
$$\Pr\left(\sup_j g'_j \leq -\frac{\epsilon z}{\sqrt{2n}}\right) \leq 0.5^M e^{-Mz^2/2}.$$

Proof of Theorem 4 (III/III)

We note that the desired inequality is trivial for $M \leq 2$. For $M \geq 3$:

$$
\begin{aligned}
&\mathbb{E} \sup_j \frac{\sqrt{2n} g'_j}{\epsilon} \\
=& -\int_{-\infty}^{0} \Pr\left(\sup_j g'_j \leq \frac{\epsilon z}{\sqrt{2n}} \right) dz + \int_0^{\infty} \Pr\left(\sup_j g'_j \geq \frac{\epsilon z}{\sqrt{2n}} \right) dz \\
\geq& \int_0^{\infty} \left(1 - (1 - 0.5 e^{-(z+1)^2/2})^M\right) dz - \int_{-\infty}^{0} 0.5^M e^{-Mz^2/2} dz \\
\geq& \int_0^{\sqrt{2 \ln M} - 1} \left(1 - 0.5(1 - 0.5/M)^M\right) dz - \int_{-\infty}^{0} 0.5^M e^{-Mz^2/2} dz \\
\geq& \int_0^{\sqrt{2 \ln M} - 1} \left(1 - 0.5/\sqrt{e}\right) dz - 0.5^{M+1} M^{-1/2} \int_{-\infty}^{\infty} e^{-z^2/2} dz \\
=& (1 - 0.5/\sqrt{e})(\sqrt{2 \ln M} - 1) - 0.5^{M+1} (3)^{-1/2} \sqrt{2\pi} \\
\geq& 0.98 \sqrt{\ln M} - 0.9.
\end{aligned}
$$

This implies the desired bound.

Majorizing Measure

A more precise characterization of Gaussian complexity, due to Talagrand 1996, is to consider a generalization of covering numbers (*majorizing measures*). We consider any measure μ on $\mathcal{F}$, and let

$$\mu(f, \epsilon, L_2(\mathcal{S}_n)) = \mu\left(\{f' \in \mathcal{F} : \|f' - f\|_{L_2(\mathcal{S}_n)} \leq \epsilon\}\right).$$

Then $-\ln \mu(f, \epsilon, L_2(\mathcal{S}_n))$ may be regarded as a generalization of the entropy number of $\mathcal{F}$ localized around f. If we define

$$\gamma_2(\mathcal{F}, \mathcal{S}_n) = \inf_{\mu} \sup_{f \in \mathcal{F}} \int_0^\infty \sqrt{\frac{-\ln \mu(f, \epsilon, L_2(\mathcal{S}_n))}{n}}\, d\epsilon.$$

Theorem 5 (The Majorizing Measure Theorem)

There exists an absolute constant $C > 0$ so that

$$C^{-1}\gamma_2(\mathcal{F}, \mathcal{S}_n) \leq G(\mathcal{F}, \mathcal{S}_n) \leq C\gamma_2(\mathcal{F}, \mathcal{S}_n).$$

Statistical Estimation

Statistical Estimation

- Observe a sample Z from a distribution $\mathcal{D}$ on $\mathcal{Z}$.
- Estimate a certain quantity $\theta \in \Theta$; based on a sample Z from

Learning Algorithm

Learning algorithm (estimator) $\mathcal{A}$: a (possibly random) map $\mathcal{Z} \to \Theta$.

Loss Function

The quality of the estimated distribution dependent quantity $\theta \in \Theta$ can be measured by a general loss function

$$Q(\theta, \mathcal{D}).$$

The goal is to find an estimator $\mathcal{A}$ that achieves the smallest loss $Q(\mathcal{A}(Z), \mathcal{D})$ when $Z \sim \mathcal{D}$.

Supervised Learning from iid Samples

The definition can handle the general setting of supervised learning, where we observe n iid training examples

$$\mathcal{S}_n = \{Z_1, \ldots, Z_n\}$$

from an unknown underlying distribution $\mathcal{D}^n$.

In this case:

- Take $Z = \mathcal{S}_n$ that is generated according to the product distribution $\mathcal{D}^n$.
- The model parameter space Θ can be regarded as the set of prediction functions.
- We may denote θ by f, so that the learning algorithm $\mathcal{A}$ learns a function $\hat{f} = \mathcal{A}(\mathcal{S}_n)$.

Example: Regression

Example 6

For least squares problem, $f(x)$ is a real valued regression function. Let

$$f_{\mathcal{D}}(x) = \mathbb{E}_{\mathcal{D}}\left[Y|X = x\right].$$

We may define

$$Q(f, \mathcal{D}) = \mathbb{E}_{X\sim\mathcal{D}}(f(X) - f_{\mathcal{D}}(X))^2.$$

Example: Conditional Density Estimation

Example 7

For conditional density estimation with K classes $y \in \{1, \ldots, K\}$, we may consider Θ as the class of vector valued density functions

$$f(x) = [p(y = 1|x), \ldots, p(y = K|x)].$$

For density estimation, the estimation quality can be measured by the KL-divergence

$$Q(f, \mathcal{D}) = \mathbb{E}_{X \sim \mathcal{D}} \mathbb{E}_{Y \sim p_{\mathcal{D}}(Y|X)} \ln \frac{p_{\mathcal{D}}(Y|X)}{p(Y|X)},$$

or by squared Hellinger distance:

$$Q(f, \mathcal{D}) = 2 - 2\mathbb{E}_{X \sim \mathcal{D}} \mathbb{E}_{Y \sim p_{\mathcal{D}}(Y|X)} \left(\frac{p(Y|X)}{p_{\mathcal{D}}(Y|X)} \right)^{1/2}.$$

Example: Classification

Example 8 (K Class Classification)

If we are interested in classification accuracy, then we may use the excess classification error over the Bayes classification error as quality measure. Here

$$f_{\mathcal{D}}(x) = \arg\max_{\ell} \; p_{\mathcal{D}}(Y = \ell | X = x)$$

is the optimal Bayes classifier. Let $f(x) \in \{1, \ldots, K\}$ be any classifier, then we can define

$$Q(f, \mathcal{D}) = \mathbb{E}_{X \sim \mathcal{D}}[\Pr(Y = f_{\mathcal{D}}(X)|X) - \Pr(Y = f(X)|X)].$$

Minimax Risk

The worst case expected risk of a learning algorithm $\mathcal{A}$ to measure the ability of the algorithm to learn the quantity θ with respect to a family of distributions $\mathcal{P}$.

Definition 9 (Def 12.9)

Consider a distribution family $\mathcal{P}$ on sample space $\mathcal{Z}$, a parameter space Θ. A learning algorithm $\mathcal{A} : \mathcal{Z}^n \to \Theta$, a loss function $Q : \Theta \times \mathcal{P} \to \mathbb{R}$. Then the worst case expected risk of a learning algorithm (i.e., a statistical estimator) $\mathcal{A}$ with respect to $\mathcal{P}$ is given by

$$r_n(\mathcal{A}, \mathcal{P}, Q) = \sup_{\mathcal{D} \in \mathcal{P}} \mathbb{E}_{S_n \sim \mathcal{D}^n} \mathbb{E}_{\mathcal{A}} \, Q(\mathcal{A}(S_n), \mathcal{D}),$$

where $\mathbb{E}_{\mathcal{A}}$ is the expectation over any internal randomization of $\mathcal{A}$. Moreover, the minimax risk is defined as:

$$r_n(\mathcal{P}, Q) = \inf_{\mathcal{A}} r_n(\mathcal{A}, \mathcal{P}, Q).$$

Minimax Analysis

In minimax analysis, we find an algorithm with the smallest worse case risk $r_n(\mathcal{A}, \mathcal{P}, Q)$.

Example 10 (Minimax Analysis for ERM)

- Upper bound of $r_n(\mathcal{P}, Q)$: can be established for specific learning algorithms. For example, if we consider the ERM method $\mathcal{A}_{\text{erm}}$ for least squares regression, then we may obtain an upper bound of

$$r_n(\mathcal{P}, Q) \leq r_n(\mathcal{A}_{\text{erm}}, \mathcal{P}, Q) = O(n^{-r}),$$

  for some $r > 0$, based on the analysis of Example 6.49.
- Lower bound: if we can show a lower bound $r_n(\mathcal{P}, Q) \geq cn^{-r}$ for some constant c that may depend on $\mathcal{P}$ but independent of n, then we know that the ERM method achieves the optimal minimax lower bound.

Lower Bound Technique: Fano's Inequality

Theorem 11 (Fano's Inequality, Thm 12.10)

Consider a finite family of distributions $\mathcal{P} = \{\mathcal{D}_1, \ldots, \mathcal{D}_N\}$. Assume that j is a random variable that is uniformly distributed in $\{1, \ldots, N\}$, and conditioned on j , $Z \sim \mathcal{D}_j$. Let $f(Z) \in \{1, \ldots, N\}$ be an estimate of the index j. Then

$$\frac{1}{N} \sum_{j=1}^{N} \Pr_{Z \sim \mathcal{D}_j} (f(Z) \neq j) \geq 1 - \frac{I(j, Z) + \ln 2}{\ln(N)},$$

where

$$I(j, Z) = \mathbb{E}_{(j,Z) \sim p(j,Z)} \ln \frac{p(j, Z)}{p(j)p(Z)}$$

is the mutual information between random variables j and Z .

Data Processing Inequality

We first state the following lemma, which says that any data process procedure (including supervised learning procedure) never increases KL divergence.

Lemma 12 (Data Processing Inequality for KL-divergence)

Consider random variables A and B, and a (possibly random) processing function h, then the inequality

$$\mathrm{KL}(A||B) \geq \mathrm{KL}(h(A)||h(B))$$

holds.

Let p_A and p_B be the densities of A and B respectively. Then the KL-divergence is defined as

$$\mathrm{KL}(A||B) = \mathbb{E}_{z\sim p_A} \ln \frac{p_A(z)}{p_B(z)}.$$

Proof of Data Processing Inequality (I/II)

Given random variables A_1, A_2 and B_1, B_2, It is easy to check that

$$\begin{aligned}
&\mathrm{KL}((A_1, A_2)||(B_1, B_2)) \\
=&\mathbb{E}_{z_1 \sim p_{A_1}} \mathbb{E}_{z_2 \sim p_{A_2|A_1}(z_2|z_1)} \ln \frac{p_{A_1}(z_1) p_{A_2}(z_2|A_1 = z_1)}{p_{B_1}(z_1) p_{B_2}(z_2|B_2 = z_1)} \\
=&\mathrm{KL}(A_1||B_1) + \mathbb{E}_{z_1 \sim p_{A_1}} \mathrm{KL}(p_{A_2}(\cdot|A_1 = z_1)||p_{B_2}(\cdot|B_1 = z_1)) \\
\geq&\mathrm{KL}(A_1||B_1).
\end{aligned}$$

Now let $A_1 = A$, $A_2 = h(A_1)$, $B_1 = B$, and $B_2 = h(B_1)$, then $P_{A_2}(\cdot|A_1 = z_1) = P_{B_2}(\cdot|B_1 = z_1)$ because the conditional probability only depends on the processing function h.

Proof of Data Processing Inequality (II/II)

Therefore

$$\mathrm{KL}(p_{A_2}(\cdot|A_1 = z_1)||p_{B_2}(\cdot|B_1 = z_1)) = 0,$$

which implies that

$$\mathrm{KL}(A||B) = \mathrm{KL}((A, h(A))||(B, h(B))).$$

Now, we change the way to evaluate the right hand side of the above equation by setting $A_2 = A$, $A_1 = h(A)$, $B_2 = B$, and $B_1 = h(B)$, which implies that

$$\mathrm{KL}((h(A), A)||(h(B), B)) \geq \mathrm{KL}(h(A)||h(B)).$$

This proves the desired result.

Lower Bound for Statistical Estimation

Theorem 13 (Generalized Fano's Inequality, Thm 12.11)

Consider a finite family of distributions $\mathcal{P} = \{\mathcal{D}_1, \ldots, \mathcal{D}_N\}$. *Given a loss function* Q *on* $\Theta \times \mathcal{P}$, *let*

$$m = \sup_{\theta \in \Theta} \left| \{k : Q(\theta, \mathcal{D}_k) < \epsilon\} \right|.$$

Assume that j *is a random variable that is uniformly distributed in* $\{1, \ldots, N\}$, *and conditioned on* j , $Z \sim \mathcal{D}_j$. *Given any (possibly random) estimator* $\mathcal{A}(Z)$. *Then*

$$\frac{1}{N} \sum_{j=1}^{N} \Pr_{Z \sim \mathcal{D}_j} (Q(\mathcal{A}(Z), \mathcal{D}_j) < \epsilon) \leq \max \left(\frac{m}{N}, \frac{I(j, Z) + \ln 2}{\ln(N/m)} \right),$$

where $I(j, Z)$ *is the mutual information of* j *and* Z. *The probability includes possible randomization in* $\mathcal{A}$.

Proof of Theorem 13 (I/III)

Let p_j be the density function of Z for $\mathcal{D}_j$. Then the joint distribution of (j, Z) is given by

$$p(j, Z) = \frac{1}{N} p_j(Z).$$

We can introduce a random variable Z' with the same marginal distribution as Z, but is independent of j:

$$p(Z') = \frac{1}{N} \sum_{j=1}^{N} p_j(Z').$$

Now, consider an arbitrary and possibly random estimator $\hat{\theta} = \mathcal{A}(Z)$. Let $\hat{\theta}' = \mathcal{A}(Z')$. By the data processing inequality for KL-divergence, with input (j, Z) and binary output $h(j, Z) = \mathbb{1}(Q(\hat{\theta}, \mathcal{D}_j) < \epsilon)$, where $\mathbb{1}(\cdot)$ is the indicator function. We obtain

$$\mathrm{KL}(\mathbb{1}(Q(\hat{\theta}, \mathcal{D}_j) < \epsilon) || \mathbb{1}(Q(\hat{\theta}', \mathcal{D}_j) < \epsilon)) \leq \mathrm{KL}((j, Z) || (j, Z')) = I(j, Z).$$

Proof of Theorem 13 (II/III)

Now let $q = \Pr(Q(\hat{\theta}, \mathcal{D}_j) < \epsilon)$ and $q' = \Pr(Q(\hat{\theta}', \mathcal{D}_j) < \epsilon)$, then the above inequality can be rewritten as:

$$\mathrm{KL}(q||q') = q \ln \frac{q}{q'} + (1-q) \ln \frac{1-q}{1-q'} \leq I(j, Z).$$

Since $\hat{\theta}'$ is independent of j, and

$$|\{j : Q(\hat{\theta}', \mathcal{D}_j) < \epsilon\}| \leq m$$

for each $\hat{\theta}'$, we obtain

$$q' \leq m/N.$$

Proof of Theorem 13 (III/III)

If $q \leq m/N$, we have proved the desired inequality. Otherwise, since $\mathrm{KL}(q||q')$ as a function of q' is decreasing in $[0, q]$, we have

$$q \ln \frac{q}{m/N} + (1-q) \ln \frac{1-q}{1-m/N} \leq I(j, Z).$$

Since $q \ln q + (1-q) \ln(1-q) \geq -\ln 2$, we obtain

$$-\ln 2 + q \ln \frac{N}{m} + (1-q) \ln \frac{N}{N-m} \leq I(j, Z).$$

This implies that

$$q \leq \frac{I(j, Z) + \ln 2}{\ln(N/m)}.$$

We thus obtain the desired bound.

Example

Example 14

In Theorem 13, if we take $\Theta = \{1, \ldots, N\}$, $Q(\theta, \mathcal{D}_j) = \mathbb{1}(\theta \neq j)$, and $\epsilon = 1$, then we have $m = 1$. Note that $1/N \leq \ln 2 / \ln(N)$, we obtain the following result

$$\frac{1}{N} \sum_{j=1}^{N} \Pr_{Z \sim \mathcal{D}_j} (\mathcal{A}(Z) = j) \leq \frac{I(j, Z) + \ln 2}{\ln N}.$$

This implies Fano's inequality of Theorem 11.

Mutual Information Bound

Lemma 15 (Lem 12.14)

The mutual information $I(j, Z)$ in Theorem 13 satisfies the inequality

$$I(j, Z) \leq \frac{1}{N^2} \sum_{j=1}^{N} \sum_{k=1}^{N} \mathrm{KL}(\mathcal{D}_j || \mathcal{D}_k) \leq \sup_{j,k} \mathrm{KL}(\mathcal{D}_j || \mathcal{D}_k).$$

The result means that if the distributions $\mathcal{D}_j$ are very similar to each other (in KL-divergence), then they are nearly independent (in mutual information).

Proof of Lemma 15

We have

$$
\begin{aligned}
I(j, Z) &= \frac{1}{N} \sum_{j=1}^{N} \mathbb{E}_{Z \sim \mathcal{D}_j} \ln \frac{p_{\mathcal{D}_j}(Z)}{\frac{1}{N} \sum_{k=1}^{N} p_{\mathcal{D}_k}(Z)} \\
&\leq \frac{1}{N^2} \sum_{j=1}^{N} \sum_{k=1}^{N} \mathbb{E}_{Z \sim \mathcal{D}_j} \ln \frac{p_{\mathcal{D}_j}(Z)}{p_{\mathcal{D}_k}(Z)},
\end{aligned}
$$

where the inequality used Jensen's inequality and the convexity of $-\ln z$.

Consequence of Mutual Information Bound

Theorem 16 (Thm 12.15)

Consider a distributions family $\mathcal{P}$ that contains a finite subset of distributions $\{\mathcal{D}_1, \ldots, \mathcal{D}_N\}$. Let Q be a loss function on $\Theta \times \mathcal{P}$, and

$$m = \sup_{\theta \in \Theta} \left| \{k : Q(\theta, \mathcal{D}_k) < \epsilon\} \right|.$$

Let $\mathcal{A}(\mathcal{S}_n)$ be an arbitrary (possibly random) estimator of $\mathcal{D}_j$ from iid data $\mathcal{S}_n = [Z_1, \ldots, Z_n] \sim \mathcal{D}_j^n$. If $m \leq N/2$ and

$$\ln(N/m) \geq \ln 4 + 2n \sup_{j,k} \mathrm{KL}(\mathcal{D}_j || \mathcal{D}_k),$$

then

$$\frac{1}{N} \sum_{j=1}^{N} \Pr_{\mathcal{S}_n \sim \mathcal{D}_j^n} \left(Q(\mathcal{A}(\mathcal{S}_n), \mathcal{D}_j) < \epsilon \right) \leq 0.5,$$

where the probability also includes possible randomization in $\mathcal{A}$. If $Q(\cdot, \cdot)$ is non-negative, then this implies that $r_n(\mathcal{P}, Q) \geq 0.5\epsilon$.

Least Squares

For the least squares regression problem in Example 6, we consider a function class $\mathcal{F}$ that contains the optimal prediction rule $f_{\mathcal{D}}(X) = \mathbb{E}[Y|X]$, and

$$Q_{\mathrm{LS}}(f, \mathcal{D}) = \mathbb{E}_{X \sim \mathcal{D}}(f(X) - f_{\mathcal{D}}(X))^2.$$

We are interested in minimax rate.

- Does ERM achieve minimax rate?
- Are there better algorithms?

Theorem 17 (Lower Bound for Least Squares, Thm 12.17)

Consider the regression model, where $X \sim \mathcal{D}_X$ with known $\mathcal{D}_X$, and

$$Y = f_{\mathcal{D}}(X) + \epsilon,$$

where ϵ is zero-mean noise that may depend on $f_{\mathcal{D}}(\cdot) \in \mathcal{F}$. Assume there exists $\sigma > 0$ so that

$$\mathbb{E}_{X \sim \mathcal{D}_X}(f(X) - f'(X))^2 \geq 2\sigma^2 \mathrm{KL}(\mathcal{D}_f || \mathcal{D}_{f'}),$$

where $\mathcal{D}_f$ is the distribution of (X, Y) when $f_{\mathcal{D}} = f$. If $\mathcal{F}$ contains N functions $f_1, \ldots, f_N$ such that

$$\ln N \geq \ln 4 + n\sigma^{-2} \sup_{j,k} \mathbb{E}_{X \sim \mathcal{D}}(f_j(X) - f_k(X))^2,$$

then

$$r_n(\mathcal{P}, Q_{\mathrm{LS}}) \geq 0.125 \inf_{j \neq k} \mathbb{E}_{X \sim \mathcal{D}_X}(f_j(X) - f_k(X))^2.$$

Proof of Theorem 17

Define $Q(f, f') = \mathbb{E}_{X \sim \mathcal{D}}(f(X) - f'(X))^2$. Note that for each $f \in \mathcal{F}$, we associate a $\mathcal{D}_f \in \mathcal{P} = \{\mathcal{D}_f : f \in \mathcal{F}\}$.

We also let $\epsilon = 0.25 \min_{j \neq k} Q(f_j, f_k)$, and it can be checked that for all $j \neq k$:

$$\max(Q(f, f_j), Q(f, f_k)) \geq (Q(f, f_j) + Q(f, f_k))/2 \geq Q(f_j, f_k)/4 \geq \epsilon.$$

This means that we can take $m = 1$, and obtain the theorem as a direct consequence of Theorem 16.

Condition of Theorem 16

Theorem 16 holds for Gaussian noise.

Proposition 18

Consider $\mathcal{D}_f(X, Y)$ so that $X \sim \mathcal{D}_X$ is identical for all $f \in \mathcal{F}$, and $Y \sim N(f(X), \sigma^2)$ for some constant $\sigma > 0$. Then

$$\mathbb{E}_{X \sim \mathcal{D}_X}(f(X) - f'(X))^2 = 2\sigma^2 \mathrm{KL}(\mathcal{D}_f || \mathcal{D}_{f'}).$$

The condition also holds for other situations, such as certain Bernoulli noise, where $Y \in \{0, 1\}$.

Metric Entropy Bound

Consider a distribution $\mathcal{D}_X$ over X, with the metric

$$\|f - f'\|_{L_2(\mathcal{D}_X)} = \left(\mathbb{E}_{X \sim \mathcal{D}_X}(f(X) - f'(X))^2\right)^{1/2}.$$

The following result shows that the corresponding metric entropy leads to a lower bound on the minimax risk.

Corollary 19 (Cor 12.20)

If for some $C > 0$ and $\epsilon > 0$:

$$C^{-1}\epsilon^{-q} \leq \ln M(\epsilon, \mathcal{F}, L_2(\mathcal{D}_X)) \leq C\epsilon^{-q}.$$

For noise model that satisfies the condition of Theorem 17, we have

$$r_n(\mathcal{P}, Q_{\mathrm{LS}}) \geq C' n^{-2/(2+q)}$$

for some $C' > 0$.

Proof of Corollary 19 (I/II)

We consider an ϵ packing subset $\mathcal{F}'$ of $\mathcal{F}$ with size of at least $\exp(C^{-1}\epsilon^{-q})$. Since for some $C_0 > 0$,

$$\ln N(0.5C_0\epsilon, \mathcal{F}, L_2(\mathcal{D}_X)) \leq 0.5C^{-1}\epsilon^{-q},$$

it implies that there exists a ball of size $0.5C_0\epsilon$, which contains at least

$$\frac{\exp(C^{-1}\epsilon^{-q})}{\exp(0.5C^{-1}\epsilon^{-q})} = \exp(0.5C^{-1}\epsilon^{-q})$$

members of $\mathcal{F}'$.
This means we can find $N \geq \exp(0.5C^{-1}\epsilon^{-q})$ functions $\{f_1, \ldots, f_N\}$ such that

$$\sup_{j\neq k} Q(f_j, f_k) \leq C_0^2\epsilon^2,$$
$$\inf_{j\neq k} Q(f_j, f_k) \geq \epsilon^2,$$

where $Q(f, f') = \mathbb{E}_{X\sim\mathcal{D}_X}(f(X) - f'(X))^2$.

Proof of Corollary 19 (II/II)

Now let $n = \lceil (C'/\epsilon^2)^{(q+2)/2} \rceil$ for a sufficiently small constant C', then we have

$$\ln N \geq 0.5 C^{-1} \epsilon^{-q} \geq \ln 4 + n\sigma^{-2} C_0^2 \epsilon^2 \geq \ln 4 + n\sigma^{-2} \sup_{j \neq k} Q(f_j, f_k).$$

Theorem 17 implies that $r_n(\mathcal{P}, Q_{\mathrm{LS}}) \geq 0.125\epsilon^2$. Since $\epsilon^2 \geq C' n^{-2/(q+2)}$, we obtain the desired bound.

ERM

Regression function class $\mathcal{F}$

- Bounded
- Has a uniform covering number of $O(\epsilon^{-q})$ for $q < 2$

We know from the local Rademacher complexity analysis of Example 6.49 that for the empirical risk minimization method

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

the following risk bound holds:

$$\mathbb{E}_{\mathcal{S}_n \sim \mathcal{D}^n} \mathbb{E}_{X \sim \mathcal{D}_X} (\hat{f}_{\text{erm}}(X) - f_{\mathcal{D}}(X))^2 = O(n^{-2/(q+2)}).$$

This matches the lower bound of Corollary 19.

ERM (cont)

For $q > 2$, the generalization bound for empirical risk minimization method is

$$\mathbb{E}_{\mathcal{S}_n \sim \mathcal{D}^n} \mathbb{E}_{X \sim \mathcal{D}_X} (\hat{f}_{\text{erm}}(X) - f_{\mathcal{D}}(X))^2 = O(n^{-1/q}),$$

and since $1/q < 2/(q+2)$, the rate is inferior to the minimax rate.

- This rate cannot be improved without additional assumptions.
- When nonparametric family has a large entropy, ERM can be suboptimal.

ERM on Sieves

It is possible to achieve the optimal rate of $O(n^{-2/(q+2)})$. One of the optimal method is least squares on sieves.

Sieve Method

Given a function class $\mathcal{F}$, instead of running least squares on $\mathcal{F}$ with

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

the sieve method considers a subset $\mathcal{F}_n \subset \mathcal{F}$, and then perform least squares regression restricted to this subset:

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

Optimal Convergence

Proposition 20 (Sieve Method Upper Bound, Prop 12.21)

Assume that the distribution of X is $\mathcal{D}_X$. Let $\mathcal{F}_n$ be an ϵ packing subset of $\mathcal{F}$ in the $L_2(\mathcal{D}_X)$ metric with $M(\epsilon, \mathcal{F}, L_2(\mathcal{D}_X))$ members. Assume there exists $b > 0$ such that $[f(X) - f'(X)] \leq 2b$ for all $f, f' \in \mathcal{F}$. Assume that $f_{\mathcal{D}} \in \mathcal{F}$ and Y is sub-Gaussian:

$$\ln \mathbb{E}_{Y|X} \exp(\lambda(Y - f_{\mathcal{D}}(X))) \leq \frac{\lambda^2 b^2}{2}.$$

Then

$$\mathbb{E}_{\mathcal{S}_n \sim \mathcal{D}^n} \mathbb{E}_{X \sim \mathcal{D}_X} (\hat{f}_{\mathcal{F}_n}(X) - f_{\mathcal{D}}(X))^2 \leq \left[4\epsilon^2 + \frac{14 b^2}{n} \ln M(\epsilon, \mathcal{F}, L_2(\mathcal{D}_X)) \right].$$

Example

Example 21

Consider the covering number condition of Corollary 19. We note that

$$\inf_{\epsilon>0} O\left(\epsilon^2 + \frac{1}{n} \ln M(\epsilon, \mathcal{F}, L_2(\mathcal{D}_X))\right) = O(n^{-2/(2+q)}).$$

Therefore the upper bound of Proposition 20 matches the lower bound of Corollary 19.

ERM Overfitting (I/II)

Consider the following function class

$$\mathcal{F} = \{f_0(x)\} \cup \{f_1(x) + \Delta f_1(x) : |\Delta f_1(x)| \leq 1/\sqrt{n}\},$$

where $f_1(x) = f_0(x) + 0.5n^{-1/4}$.

If we consider ERM with sieve $\mathcal{F}_n = \{f_k(x) : k = 0, 1\}$, then we have a convergence rate no worse than $O(1/n)$.

However, we have a overfitting problem with ERM on $\mathcal{F}$. To see this, we may consider the model

$$Y = f_0(X) + \epsilon, \quad \epsilon \sim N(0, 1),$$

and training data $(X_1, Y_1), \ldots, (X_n, Y_n)$.

ERM Overfitting (II/II)

Let $\delta_i = \Delta f_1(X_i)$ and $\epsilon_i = Y_i - f_0(X_i)$. Then we have

$$\sum_{i=1}^{n} \min_{|\delta_i| \le 1/\sqrt{n}} (f_1(X_i) + \delta_i - Y_i)^2 - \sum_{i=1}^{n} (f_0(X_i) - Y_i)^2$$

$$= \sum_{i=1}^{n} \min_{|\delta_i| \le 1/\sqrt{n}} (0.5n^{-1/4} + \delta_i - \epsilon_i)^2 - \sum_{i=1}^{n} \epsilon_i^2$$

$$\le \sqrt{n}(0.5 + n^{-1/4})^2 - n^{-1/4} \sum_{i=1}^{n} \epsilon_i - 2n^{-1/2} \sum_{i=1}^{n} |\epsilon_i|.$$

The inequality is achieved with $\delta_i = \text{sign}(\epsilon_i)/\sqrt{n}$. When n is large,

$$-2n^{-1/2} \sum_{i=1}^{n} |\epsilon_i| = -2n^{1/2} \left[\mathbb{E}_{\epsilon \sim N(0,1)} |\epsilon| + O_p(1/\sqrt{n}) \right]$$

dominates the RHS with large probability. It implies that with large probability, ERM gives an estimator $\hat{f}(x) = f_1(x) + \Delta f_1(x)$ with $|\Delta f(x)| \le 1/\sqrt{n}$. This means that

$$\mathbb{E}_{\mathcal{S}_n} \mathbb{E}_X (\hat{f}(X) - f(X))^2 \ge c/\sqrt{n}$$

for some $c > 0$. Note that this is a suboptimal rate.

Lower Bound: Assouad's Lemma

Lemma 22 (Assouad's Lemma, Simplified Lem 12.27)

Let $d \geq 1$ be an integer and $\mathcal{P}_d = \{\mathcal{D}_\tau : \tau \in \{-1,1\}^d\}$ contain 2^d probability measures. Suppose that the loss function Q can be decomposed as $Q(\theta, \mathcal{D}) = \sum_{j=1}^d Q_j(\theta, \mathcal{D})$. For any j and τ, let $\tau^{-[j]}$ be the index that differs with τ only by one coordinate j. Assume that there exists $\epsilon, \beta_j \geq 0$ such that

$$\forall \tau : \quad [Q_j(\theta, \mathcal{D}_\tau) + Q_j(\theta, \mathcal{D}_{\tau^{-[j]}})] \geq \epsilon, \quad \|\mathcal{D}_\tau - \mathcal{D}_{\tau^{-[j]}}\|_{\mathrm{TV}} \leq \beta_j.$$

Consider randomized $\mathcal{A}(\mathcal{S}_n)$ based on $\mathcal{S}_n \sim \mathcal{D}_\tau^n$ for some τ. We have

$$\frac{1}{2^d} \sum_\tau \mathbb{E}_{\mathcal{S}_n \sim \mathcal{D}_\tau^n} \mathbb{E}_{\mathcal{A}} Q(\mathcal{A}(\mathcal{S}_n), \mathcal{D}_\tau) \geq \frac{\epsilon d}{2} - \frac{\epsilon}{2} \sum_{j=1}^d \beta_j,$$

where $\mathbb{E}_{\mathcal{A}}$ is with respect to the internal randomization in $\mathcal{A}$. This implies that

$$r_n(\mathcal{P}_d, Q) \geq \frac{\epsilon d}{2} - \frac{\epsilon}{2} \sum_{j=1}^d \beta_j.$$

Proof of Lemma 22

For notation convenient, assume there exists $d\mu(z)$ so that for all $\mathcal{D} \in \mathcal{P}$, $p_{\mathcal{D}}(z)d\mu(z)$ is the distribution of $\mathcal{D}$. Then

$$\frac{2}{|\mathcal{P}|}\sum_{j=1}^{d}\sum_{\mathcal{D}\in\mathcal{P}}\mathbb{E}_{Z\sim\mathcal{D}}Q_j(\mathcal{A}(Z),\mathcal{D})$$

$$=\frac{1}{|\mathcal{P}|}\sum_{j=1}^{d}\mathbb{E}_{Z\sim\mathcal{D}}\sum_{\tau}[Q_j(\mathcal{A}(Z),\mathcal{D}_\tau)+Q_j(\theta,\mathcal{D}_{\tau-[j]})]$$

$$=\frac{1}{|\mathcal{P}|}\sum_{j=1}^{d}\sum_{\tau}\int[Q_j(\mathcal{A}(z),\mathcal{D}_\tau)p_{\mathcal{D}_\tau}(z)+Q_j(\mathcal{A}(z),\mathcal{D}_{\tau-[j]})p_{\mathcal{D}_{\tau-[j]}}(z)]d\mu(z)$$

$$\geq\frac{1}{|\mathcal{P}|}\sum_{j=1}^{d}\sum_{\tau}\int\epsilon\min(p_{\mathcal{D}_\tau}(z),p_{\mathcal{D}_{\tau-[j]}}(z))d\mu(z)$$

(Lemma's assumption)

$$=\frac{1}{|\mathcal{P}|}\sum_{j=1}^{d}\sum_{\tau}\epsilon(1-\|\mathcal{D}_\tau-\mathcal{D}_{\tau-[j]}\|_{\mathrm{TV}})\geq\sum_{j=1}^{d}\epsilon(1-\beta_j).$$ (TV-norm def)

Theorem 23 (Thm 12.28)

Under the assumptions of Lemma 22. For any j and τ, let $\tau^{-[j]}$ be the index that differs with τ only by one coordinate j. Assume that there exists $\epsilon, \beta_j \geq 0$ such that

$$\forall \tau: \quad [Q_j(\theta, \mathcal{D}_\tau) + Q_j(\theta, \mathcal{D}_{\tau^{-[j]}})] \geq \epsilon, \quad H(\mathcal{D}_\tau || \mathcal{D}_{\tau^{-[j]}}) \leq \beta_j.$$

Consider randomized $\mathcal{A}(\mathcal{S}_n)$ based on $\mathcal{S}_n \sim \mathcal{D}_\tau^n$ for some τ. We have

$$\frac{1}{2^d} \sum_\tau \mathbb{E}_{\mathcal{S}_n \sim \mathcal{D}_\tau^n} \mathbb{E}_{\mathcal{A}} Q(\mathcal{A}(\mathcal{S}_n), \mathcal{D}_\tau) \geq \frac{\epsilon d}{2} - \frac{\epsilon}{2} \sum_{j=1}^{d} \sqrt{2 - 2(1 - 0.5\beta_j^2)^n},$$

where $\mathbb{E}_{\mathcal{A}}$ is with respect to the internal randomization in $\mathcal{A}$. This implies that

$$r_n(\mathcal{P}_d, Q) \geq \frac{\epsilon d}{2} - \frac{\epsilon}{2} \sum_{j=1}^{d} \sqrt{2 - 2(1 - 0.5\beta_j^2)^n}.$$

Example of Assouad's Lemma

Consider observations $Z_i \in \{0,1\}^d$, where each Z_i has d components $Z_{i,j} \sim \text{Bernoulli}(\theta_j)$ for $j = 1, \ldots, d$.

Let $\theta = [\theta_1, \ldots, \theta_d] \in (0,1)^d$ be the model parameters to be estimated. For $\tau \in \{\pm 1\}^d$, we let $\theta_{\tau,j} = \epsilon^2(1+\tau_j)/2$, where $\epsilon \in (0, 0.5)$. Let $\mathcal{D}_\tau$ be the corresponding Bernoulli distribution, and $\mathcal{P}_d = \{\mathcal{D}_\tau\}$. Define the metric

$$Q(\hat{\theta}, \theta) = \sum_{j=1}^{d} Q_j(\hat{\theta}, \theta), \quad Q_j(\hat{\theta}, \theta) = \left| \sqrt{\hat{\theta}_j} - \sqrt{\theta_j} \right|.$$

We cannot apply Theorem 16 directly on this subclass $\mathcal{P}_d$ because the KL-divergence of two distributions in $\mathcal{P}_d$ can be infinity.

Example (cont)

On the other hand, for all τ:

$$[Q_j(\theta, \mathcal{D}_\tau) + Q_j(\theta, \mathcal{D}_{\tau - \lfloor j \rfloor})] \geq \epsilon$$

and

$$H(\mathcal{D}_\tau || \mathcal{D}_{\tau - \lfloor j \rfloor}) \leq 2\epsilon.$$

We thus obtain from Theorem 23 that

$$r_n(\mathcal{P}_d, Q) \geq 0.5d\epsilon - 0.5d\epsilon\sqrt{2 - 2(1 - 2\epsilon^2)^n}.$$

For sufficiently small ϵ, with $n \leq 1/(6\epsilon^2)$, we obtain

$$r_n(\mathcal{P}_d, Q) \geq 0.1d\epsilon.$$

Summary (Chapter 12)

- Gaussian complexity versus Rademacher complexity
  - Slepian's lemma
- Empirical processes and covering numbers
  - upper bound using chaining
  - lower bound using Sudakov minoration
- Gaussian complexity and Majorizing Measures
  - covering number $\rightarrow$ Majorizing measures
  - chaining $\rightarrow$ generic chaining
  - tight bounds
- Statistical estimation
- Data processing inequality
- Lower Bound for Statistical Estimation
- Entropy based Minimax Rate for Least Squares
- Matching upper bound: ERM on sieves
- Overfitting of ERM
- Lower Bound from Assouad's lemma