Uniform Convergence

Mathematical Analysis of Machine Learning Algorithms
(Chapter 3)

PAC Learning: Notations

We will study Probabilistic Approximately Correct (PAC) learning.

- X: binary valued vector $X \in \{0, 1\}^d$.
- Y: binary output $Y \in \{0, 1\}$.
- f: a Boolean function: $X \in \{0, 1\}^d \to Y \in \{0, 1\}$.
- $\mathcal{C}$: a set of Boolean functions
- $f_* \in \mathcal{C}$: unknown true function that we want to learn.
- $\mathcal{O}$: an oracle that sample from a distribution $\mathcal{D}$, each sample return $X \sim \mathcal{D}$ and $Y = f(X_*)$

The goal of a PAC learner is to learn $f_*(X)$ so that generalization error

$$\mathrm{err}_{\mathcal{D}}(f) = \mathbb{E}_{X \sim \mathcal{D}} \mathbb{1}(f(x) \neq f_*(x)).$$

is no larger than ϵ.

PAC Learning: Definition

We may call the oracle $\mathcal{O}$ n times to form a training data $\mathcal{S}_n = \{(X_i, Y_i)\}_{i=1,\ldots,n} \sim \mathcal{D}^n$. The learner $\mathcal{A}$ takes $\mathcal{S}_n$ and returns a function $\hat{f} \in \mathcal{C}$.

Definition 1 (PAC Learning)

A concept class $\mathcal{C}$ is PAC learnable if there exists a learner $\mathcal{A}$ so that for all $f_* \in \mathcal{C}$, distribution $\mathcal{D}$ on the input, approximation error $\epsilon > 0$ and probability $\delta \in (0, 1)$, the following statement holds. With probability at least $1 - \delta$ over samples from the oracle $\mathcal{O}$ over $\mathcal{D}$, the learner produces a function $\hat{f}$ such that

$$\mathrm{err}_{\mathcal{D}}(\hat{f}) \leq \epsilon,$$

with the computational complexity polynomial in $(\epsilon^{-1}, \delta^{-1}, d)$.

In the statistical complexity analysis of learning algorithms, the computational complexity requirement is de-emphasized.

PAC Learning: examples

Example 2 (AND Function Class)

Each member of **AND** function class can be written as

$$f(x) = \prod_{j \in J} x_j, \qquad J \subset \{1, \ldots, d\}.$$

Example 3 (Decision List)

A decision list is a function of the following form. Let $\{i_1, \ldots, i_d\}$ be a permutation of $\{1, \ldots, d\}$, and let $a_i, b_i \in \{0, 1\}$ for $i = 1, \ldots, d+1$. The function $f(x)$ can be computed as follows. if $x_{i_1} = a_1$ then $f(x) = b_1$; else if $x_{i_2} = a_2$ then $f(x) = b_2$, $\cdots$, else if $x_{i_d} = a_d$ then $f(x) = b_d$; else $f(x) = b_{d+1}$.

ERM

Definition 4 (ERM)

Define the training error of $f \in \mathcal{C}$ as

$$\widehat{\mathrm{err}}_{\mathcal{S}_n}(f) = \frac{1}{n} \sum_{i=1}^{n} \mathbb{1}(f(X_i) \neq Y_i).$$

The ERM (empirical risk minimization) method finds a function $\hat{f} \in \mathcal{C}$ that minimizes the training error.

Since by the realizable assumption of PAC learning, $f_* \in \mathcal{C}$ achieves zero training error, the empirical minimizer $\hat{f}$ that achieves zero training error. More generally, we may consider approximate ERM, which returns $\hat{f}$ so that

$$\widehat{\mathrm{err}}_{\mathcal{S}_n}(\hat{f}) \leq \epsilon' \tag{1}$$

for some accuracy $\epsilon' > 0$.

Analysis of PAC Learning: Decomposition

We want to estimate the difference of the test error $\mathrm{err}_{\mathcal{D}}(\hat{f})$ and the optimal test error $\mathrm{err}_{\mathcal{D}}(f_*)$:

$$
\begin{aligned}
&\mathrm{err}_{\mathcal{D}}(\hat{f}) - \mathrm{err}_{\mathcal{D}}(f_*) \\
=&\underbrace{[\mathrm{err}_{\mathcal{D}}(\hat{f}) - \widehat{\mathrm{err}}_{\mathcal{S}_n}(\hat{f})]}_{A} + \underbrace{[\widehat{\mathrm{err}}_{\mathcal{S}_n}(\hat{f}) - \widehat{\mathrm{err}}_{\mathcal{S}_n}(f_*)]}_{B} + \underbrace{[\widehat{\mathrm{err}}_{\mathcal{S}_n}(f_*) - \mathrm{err}_{\mathcal{D}}(f_*)]}_{C} \\
\leq&\underbrace{\sup_{f\in\mathcal{F}}[\mathrm{err}_{\mathcal{D}}(f) - \widehat{\mathrm{err}}_{\mathcal{S}_n}(f)]}_{A'} + 0 + \underbrace{[\widehat{\mathrm{err}}_{\mathcal{S}_n}(f_*) - \mathrm{err}_{\mathcal{D}}(f_*)]}_{C} \\
\leq&2\underbrace{\sup_{f\in\mathcal{F}}\left|\mathrm{err}_{\mathcal{D}}(f) - \widehat{\mathrm{err}}_{\mathcal{S}_n}(f)\right|}_{A''}.
\end{aligned}
$$

The quantity A' or A'' requires that the convergence of empirical mean to the true mean holds for all $f \in \mathcal{F}$.
Such a convergence result is referred to as *uniform convergence*.

Analysis of PAC Learning: Union Bound

The key mathematical tool to analyze uniform convergence is the *union bound*, described in Proposition 5.

Proposition 5 (Union Bound)

Consider m events $E_1, \ldots E_m$. The following probability inequality holds:

$$\Pr\left(E_1 \cup \cdots \cup E_m\right) \leq \sum_{j=1}^{m} \Pr(E_j).$$

Alternative Expression of Union Bound

Assume each event E_j occurs with probability at least $1 - \delta_j$ for $j = 1, \ldots, m$, then with probability at least $1 - \sum_{j=1}^{m} \delta_j$:

All of events $\{E_j\}$ occur simultaneously for $j = 1, \ldots, m$.

Uniform Convergence Analysis

We apply the additive Chernoff bound to obtain for each fixed $f \in \mathcal{C}$:

$$\Pr\left(\mathrm{err}_{\mathcal{D}}(f) \geq \widehat{\mathrm{err}}_{\mathcal{S}_n}(f) + \epsilon\right) \leq \exp(-2n\epsilon^2).$$

Remarks:

- We cannot directly apply the Chernoff bound to the function $\hat{f}$ learned from the training data $\mathcal{S}_n$, because $\hat{f}$ is a random function that depends on $\mathcal{S}_n$.
- We need union bound to handle $\hat{f}$, which we will demonstrate next.

Uniform Convergence Analysis: union bound

We can now take the union bound as follows:

$$\begin{aligned}
&\Pr\left(\sup_{f\in\mathcal{C}}\left[\mathrm{err}_{\mathcal{D}}(f)-\widehat{\mathrm{err}}_{\mathcal{S}_n}(f)\right]\geq\epsilon\right)\\
=&\Pr\left(\exists f\in\mathcal{C}:\mathrm{err}_{\mathcal{D}}(f)\geq\widehat{\mathrm{err}}_{\mathcal{S}_n}(f)+\epsilon\right)\\
\leq&\sum_{f\in\mathcal{C}}\Pr\left(\mathrm{err}_{\mathcal{D}}(f)\geq\widehat{\mathrm{err}}_{\mathcal{S}_n}(f)+\epsilon\right)\\
\leq&N\exp(-2n\epsilon^2).
\end{aligned}$$

Such a result (which implies that with large probability, error is small for all $f\in\mathcal{C}$) is called *uniform convergence*.

Uniform Convergence Analysis: alternative expression

Now by setting $N \exp(-2n\epsilon^2) = \delta$ and solving for ϵ to get

$$\epsilon = \sqrt{\frac{\ln(N/\delta)}{2n}},$$

we obtain the following equivalent statement.

Uniform Convergence for Finite $\mathcal{C}$

With probability at least $1 - \delta$, the following inequality holds for all $f \in \mathcal{C}$:

$$\mathrm{err}_{\mathcal{D}}(f) < \widehat{\mathrm{err}}_{\mathcal{S}_n}(f) + \sqrt{\frac{\ln(N/\delta)}{2n}}.$$

Consequence of Uniform Convergence

Given sample $\mathcal{S}_n$, a uniform convergence bound holds for all $f \in \mathcal{C}$. Therefore it holds for the output $\hat{f} \in \mathcal{C}$ from any learning algorithm.

Oracle Inequality

Oracle Inequality

With probability at least $1-\delta$, the following inequality holds for the ERM PAC learner (1) for all $\gamma > 0$:

$$\mathrm{err}_{\mathcal{D}}(\hat{f}) < \epsilon' + \sqrt{\frac{\ln(N/\delta)}{2n}} = (1+\gamma)\sqrt{\frac{\ln(N/\delta)}{2n}}, \tag{2}$$

with

$$\epsilon' = \gamma\sqrt{\frac{\ln(N/\delta)}{2n}}.$$

It can be expressed in another form of sample complexity bound. If we let

$$n \geq \frac{(1+\gamma)^2 \ln(N/\delta)}{2\epsilon^2},$$

then $\mathrm{err}_{\mathcal{D}}(\hat{f}) < \epsilon$ with probability at least $1-\delta$.

Better Generalization Bound

Theorem 6 (Thm 3.6)

Consider a concept class $\mathcal{C}$ with N elements. With probability at least $1 - \delta$*, the ERM PAC learner* (1) *with*

$$\epsilon' = \gamma^2 \frac{2 \ln(N/\delta)}{n}$$

for some $\gamma > 0$ *satisfies*

$$\mathrm{err}_{\mathcal{D}}(\hat{f}) \leq (1 + \gamma)^2 \frac{2 \ln(N/\delta)}{n}.$$

Sample Complexity

Theorem 6 is stated in statistical convergence of $O(1/n)$ rate.
It implies the following equivalent sample complexity bound.

Sample Complexity Bound

Given $\delta \in (0, 1)$. For all sample size

$$n \geq (1 + \gamma)^2 \frac{2 \ln(N/\delta)}{\epsilon},$$

we have with probability at least $1 - \delta$:

$$\mathrm{err}(\hat{f}) < \epsilon.$$

Example

Example 7

The AND concept class $\mathcal{C}$ is PAC learnable. To show this, we will prove that the ERM (1) solution can be obtained in a computationally efficient way with $\epsilon' = 0$. If this is true, then Theorem 6 implies that $\mathcal{C}$ is PAC-learnable because the number of AND functions cannot be more than $N = 2^d$. Therefore $\ln N \leq d \ln 2$.
In the following, we show that ERM solution can be efficiently obtained. Given $\mathcal{S}_n = \{(X_1, Y_1), \ldots, (X_n, Y_n)\} \sim \mathcal{D}^n$, we define $\hat{J} = \{j : \quad \forall 1 \leq i \leq n, X_{i,j} \geq Y_i\}$ (where X_{ij} denotes the j-th component of the i-th training data X_i) and $\hat{f}(x) = \prod_{j \in \hat{J}} x_j$. This choice implies that $\hat{f}(X_i) = Y_i$ when $Y_i = 1$. It can be easily verified that if the true target is $f_*(x) = \prod_{j \in J} x_j$, then $\hat{J} \supset J$. This implies that $\hat{f}(x) \leq f_*(x)$. This implies that $\hat{f}(X_i) = Y_i$ when $Y_i = 0$, and hence $\widehat{\mathrm{err}}_{\mathcal{S}_n}(\hat{f}) = 0$.

Proof of Theorem 6 (I/II)

Given any $f \in \mathcal{C}$, we have from Corollary 2.18 that

$$\Pr\left(\mathrm{err}_{\mathcal{D}}(f) \geq \widehat{\mathrm{err}}_{\mathcal{S}_n}(f) + \epsilon\right) \leq \exp\left(\frac{-n\epsilon^2}{2\mathrm{err}_{\mathcal{D}}(f)}\right).$$

Now by setting $\exp(-n\epsilon^2/2\mathrm{err}_{\mathcal{D}}(f)) = \delta/N$, and solve for ϵ:

$$\epsilon = \sqrt{\frac{2\mathrm{err}_{\mathcal{D}}(f)\ln(N/\delta)}{n}},$$

we obtain the following equivalent statement. With probability at least $1 - \delta/N$:

$$\mathrm{err}_{\mathcal{D}}(f) \leq \widehat{\mathrm{err}}_{\mathcal{S}_n}(f) + \sqrt{\frac{2\mathrm{err}_{\mathcal{D}}(f)\ln(N/\delta)}{n}}.$$

Proof of Theorem 6 (II/II)

The union bound thus implies the following statement. With probability at least $1 - \delta$, for all $f \in \mathcal{C}$:

$$\mathrm{err}_{\mathcal{D}}(f) \leq \widehat{\mathrm{err}}_{\mathcal{S}_n}(f) + \sqrt{\frac{2\mathrm{err}_{\mathcal{D}}(f) \ln(N/\delta)}{n}}.$$

The inequality also holds for the ERM PAC learner solution (1). Thus

$$\begin{aligned}\mathrm{err}_{\mathcal{D}}(\hat{f}) \leq& \widehat{\mathrm{err}}_{\mathcal{S}_n}(\hat{f}) + \sqrt{\frac{2\mathrm{err}_{\mathcal{D}}(\hat{f}) \ln(N/\delta)}{n}} \\ \leq& \gamma^2 \frac{2\ln(N/\delta)}{n} + \sqrt{\frac{2\mathrm{err}_{\mathcal{D}}(\hat{f}) \ln(N/\delta)}{n}}.\end{aligned}$$

We can solve the above inequality for $\mathrm{err}_{\mathcal{D}}(\hat{f})$ and obtain

$$\mathrm{err}_{\mathcal{D}}(\hat{f}) \leq (\gamma^2 + 0.5 + \sqrt{\gamma^2 + 0.25}) \frac{2\ln(N/\delta)}{n},$$

which implies the desired bound as $\gamma^2 + 0.5 + \sqrt{\gamma^2 + 0.25} \leq (1+\gamma)^2$.

Empirical Process

The analysis of realizable PAC learning can be generalized to deal with

- general non-binary-valued functions
- functions classes which may contain an infinitely number of functions
- handle the non-realizable case where $f_*(x) \notin C$ or when the observation Y contains noise.

For such cases, the corresponding analysis requires the technical tool of empirical processes.

Notations

To simplify the notations, in the general setting, we may denote

- Observations as $Z_i = (X_i, Y_i) \in \mathcal{Z} = \mathcal{X} \times \mathcal{Y}$
- Loss function as $L(f(X_i), Y_i)$.
- Prediction function as $f(X_i)$ (which is often a vector-valued-function)
- Assume further that $f(x)$ is parametrized by $w \in \Omega$ as $f(w, x)$
- Hypothesis space is $\{f(w, \cdot) : w \in \Omega\}$.
- Training data $\mathcal{S}_n = \{Z_i = (X_i, Y_i) : i = 1, \ldots, n\}$.

Notations Simplified

Definition 8

We define

$$\phi(w,z) = L(f(w,x),y) - L_*(x,y), \tag{3}$$

for $w \in \Omega$ and $z = (x,y) \in \mathcal{Z} = \mathcal{X} \times \mathcal{Y}$, and a pre-chosen $L_*(x,y)$ of $z = (x,y)$ that does not depend on w. Define

$$\phi(w,\mathcal{S}_n) = \frac{1}{n}\sum_{i=1}^{n} \phi(w,Z_i). \tag{4}$$

Moreover, for a distribution $\mathcal{D}$ on $\mathcal{Z}$, we define the test loss for $w \in \Omega$

$$\phi(w,\mathcal{D}) = \mathbb{E}_{Z\in\mathcal{D}}\phi(w,Z). \tag{5}$$

In many cases, we can set $L_*(x,y) = 0$.

Example of Nonzero $L_*(x, y)$

Although we can set $L_*(x, y) = 0$, we can also choose it so that $\phi(w, z)$ has a small variance.

Example 9

Consider linear model $f(w, x) = w^\top x$, and let $L(f(w, x), y) = (w^\top x - y)^2$ be the least squares loss. Then with $L_*(x, y) = 0$, we have $\phi(w, z) = (w^\top x - y)^2$ for $z = (x, y)$.

If we further assume that the problem is realizable by linear model, and w_* is the true weight vector: $\mathbb{E}[y|x] = w_*^\top x$. It follows that we may take $L_*(x, y) = (w_*^\top x - y)^2$, and

$$\phi(w, z) = (w^\top x - y)^2 - (w_*^\top x - y)^2,$$

which has a small variance when $w \approx w_*$ because $\lim_{w \to w_*} \phi(w, z) = 0$.

Uniform Convergence

Definition 10 (Uniform Convergence)

Given a model space Ω, and distribution $\mathcal{D}$. Let $\mathcal{S}_n \sim \mathcal{D}^n$ be n iid examples sampled from $\mathcal{D}$ on $\mathcal{Z}$. We say that $\phi(w, \mathcal{S}_n)$ $(w \in \Omega)$ converges to $\phi(w, \mathcal{D})$ uniformly in probability if for all $\epsilon > 0$:

$$\lim_{n \to \infty} \Pr\left(\sup_{w \in \Omega} |\phi(w, \mathcal{S}_n) - \phi(w, \mathcal{D})| > \epsilon \right) = 0,$$

where the probability is over iid samples of $\mathcal{S}_n \sim \mathcal{D}^n$.

Approximate ERM

We consider a more general form of ERM, approximate ERM, which satisfies the following inequality for some $\epsilon' > 0$:

$$\frac{1}{n}\sum_{i=1}^{n} L(f(\hat{w}, X_i), Y_i) \leq \inf_{w \in \Omega} \left[\frac{1}{n}\sum_{i=1}^{n} L(f(w, X_i), Y_i)\right] + \epsilon'. \tag{6}$$

The quantity $\epsilon' > 0$ indicates how accurately we solve the ERM problem.
Since L_* is independent of w, the approximate ERM method (6) becomes

$$\phi(\hat{w}, \mathcal{S}_n) \leq \inf_{w \in \Omega} \phi(w, \mathcal{S}_n) + \epsilon' \tag{7}$$

in our notation.

Uniform Convergence Implies Oracle Inequality

Lemma 11 (Simplification of Lem 3.11)

Assume that for any $\delta \in (0, 1)$*, the following uniform convergence result holds. With probability at least* $1 - \delta_1$*,*

$$\forall w \in \Omega : \; \phi(w, \mathcal{D}) \le \phi(w, \mathcal{S}_n) + \epsilon_n(\delta_1, w).$$

Moreover, $\forall w \in \Omega$*, the following inequality holds. With probability at least* $1 - \delta_2$*,*

$$\phi(w, \mathcal{S}_n) \le \phi(w, \mathcal{D}) + \epsilon'_n(\delta_2, w).$$

Then the following statement holds. With probability at least $1 - \delta_1 - \delta_2$*, the approximate ERM method* (7) *satisfies the oracle inequality:*

$$\phi(\hat{w}, \mathcal{D}) \le \inf_{w \in \Omega} \left[\phi(w, \mathcal{D}) + \epsilon'_n(\delta_2, w)\right] + \epsilon' + \epsilon_n(\delta_1, \hat{w}).$$

A more general version is presented in the book.

Proof of Lemma 11

Consider an arbitrary $w \in \Omega$. We have with probability at least $1 - \delta_1$:

$$\begin{aligned} \phi(\hat{w}, \mathcal{D}) \leq& \phi(\hat{w}, \mathcal{S}_n) + \epsilon_n(\delta_1, \hat{w}) \\ \leq& \phi(w, \mathcal{S}_n) + \epsilon' + \epsilon_n(\delta_1, \hat{w}). \end{aligned} \tag{8}$$

Moreover, with probability at least $1 - \delta_2$:

$$\phi(w, \mathcal{S}_n) \leq \phi(w, \mathcal{D}) + \epsilon'_n(\delta_2, w). \tag{9}$$

Taking the union bound of the two events, we obtain with probability at least $1 - \delta_1 - \delta_2$, both (8) and (9) hold. It follows that

$$\begin{aligned} \phi(\hat{w}, \mathcal{D}) \leq& \phi(w, \mathcal{S}_n) + \epsilon' + \epsilon_n(\delta_1, \hat{w}) \\ \leq& \phi(w, \mathcal{D}) + \epsilon'_n(\delta_2, w) + \epsilon' + \epsilon_n(\delta_1, \hat{w}). \end{aligned}$$

Since w is arbitrary, we let w approach the minimum of the right hand side, and obtain the desired bound.

Covering Number (Bracketing Number)

If Ω is finite, then we can use union bound to obtain uniform convergence of empirical processes. If Ω is infinite, then we can approximate the function class

$$\mathcal{G} = \{\phi(w, z) : w \in \Omega\}$$

using a finite function class.

Definition 12 (Lower Bracketing Cover)

Given a distribution $\mathcal{D}$. A finite function class $\mathcal{G}(\epsilon) = \{\phi_1(z), \ldots, \phi_N(z)\}$ is an ϵ lower bracketing cover of $\mathcal{G}$ (with $L_1(\mathcal{D})$ metric) if for all $w \in \Omega$, there exists $j = j(w)$ such that

$$\forall z : \phi_j(z) \leq \phi(w, z), \qquad \mathbb{E}_{Z\sim\mathcal{D}}\phi_j(Z) \geq \mathbb{E}_{Z\sim\mathcal{D}}\phi(w, Z) - \epsilon.$$

The ϵ-lower bracketing number of $\mathcal{G}$, denoted by $N_{LB}(\epsilon, \mathcal{G}, L_1(\mathcal{D}))$, is the smallest cardinality of such $\mathcal{G}(\epsilon)$. The quantity $\ln N_{LB}(\epsilon, \mathcal{G}, L_1(\mathcal{D}))$ is referred to as the ϵ-lower bracketing entropy.

The functions $\phi_j(z)$ may not necessarily belong to $\mathcal{G}$.

Uniform Convergence Analysis

Theorem 13 (Simplification of Thm 3.14)

Assume that $\phi(w, z) \in [0, 1]$ *for all* $w \in \Omega$ *and* $z \in \mathcal{Z}$. *Let* $\mathcal{G} = \{\phi(w, z) : w \in \Omega\}$. *Then given* $\delta \in (0, 1)$, *with probability at least* $1 - \delta$, *the following inequality holds:*

$$\forall w \in \Omega : \ \phi(w, \mathcal{D}) \leq [\phi(w, \mathcal{S}_n) + \epsilon_n(\delta, \mathcal{G}, \mathcal{D})],$$

where

$$\epsilon_n(\delta, \mathcal{G}, \mathcal{D}) = \inf_{\epsilon > 0} \left[\epsilon + \sqrt{\frac{\ln(N_{LB}(\epsilon, \mathcal{G}, L_1(\mathcal{D}))/\delta)}{2n}} \right].$$

This result employs additive Chernoff bound. There is also a version using multiplicative Chernoff bound (see Theorem 3.14).

Proof of Theorem 13 (I/II)

For any $\epsilon > 0$, let $\mathcal{G}(\epsilon) = \{\phi_1(z), \ldots, \phi_N(z)\}$ be an ϵ lower bracketing cover of $\mathcal{G}$ with $N = N_{LB}(\epsilon, \mathcal{G}, L_1(\mathcal{D}))$.
We may assume that $\phi_j(z) \in [0, 1]$ for all j because otherwise, we may set $\phi_j(z)$ to

$$\min(1, \max(0, \phi_j(z))).$$

In the following, we let $j = j(w)$ for simplified notation:

$$\begin{aligned} &\frac{1}{n}\sum_{i=1}^{n} \phi(w, Z_i) - \mathbb{E}_{Z\sim\mathcal{D}}\phi(w, Z) \\ \geq &\frac{1}{n}\sum_{i=1}^{n} \phi_j(Z_i) - \mathbb{E}_{Z\sim\mathcal{D}}\phi_j(Z) - \epsilon. \end{aligned} \tag{10}$$

Proof of Theorem 13 (II/II)

Let $\epsilon'' = \sqrt{\ln(N/\delta)/2n}$. It follows from the union bound on j that

$$\begin{aligned}
&\Pr\left(\exists w \in \Omega : \left[\frac{1}{n}\sum_{i=1}^{n}\phi(w, Z_i) - \mathbb{E}_{Z\sim\mathcal{D}}\phi(w, Z) + \epsilon + \epsilon''\right] \leq 0\right) \\
\leq &\Pr\left(\exists j : \left[\frac{1}{n}\sum_{i=1}^{n}\phi_j(Z_i) - \mathbb{E}_{Z\sim\mathcal{D}}\phi_j(Z) + \epsilon''\right] \leq 0\right) \\
\leq &\sum_{j=1}^{N}\Pr\left(\frac{1}{n}\sum_{i=1}^{n}\phi_j(Z_i) - \mathbb{E}_{Z\sim\mathcal{D}}\phi_j(Z) + \epsilon'' \leq 0\right) \\
\leq &N\exp(-2n(\epsilon'')^2) = \delta.
\end{aligned}$$

This implies the desired bound.

Generalization (Oracle Inequality)

The uniform convergence bounds in Theorem 13 imply generalization bounds as follows.

Corollary 14 (Simplification of Cor 3.15)

Assume that $\phi(w, z) \in [0, 1]$ *for all* $w \in \Omega$ *and* $z \in \mathcal{Z}$*. Let* $\mathcal{G} = \{\phi(w, z) : w \in \Omega\}$*. With probability at least* $1 - \delta$*, the approximate ERM method* (7) *satisfies the (additive) oracle inequality:*

$$\phi(\hat{w}, \mathcal{D}) \leq \inf_{w \in \Omega} \phi(w, \mathcal{D}) + \epsilon' + \inf_{\epsilon > 0} \left[\epsilon + \sqrt{\frac{2 \ln(2 N_{LB}(\epsilon, \mathcal{G}, L_1(\mathcal{D}))/\delta)}{n}} \right].$$

We may take $\phi(w, z) = L(f(w, x), y)$ with $L_*(x, y) = 0$ to obtain an oracle inequality for the approximate ERM method (6).

Proof

We can take $\epsilon_n(\delta/2, w) = \epsilon_n(\delta/2, \mathcal{G}, \mathcal{D})$, as defined in Theorem 13.
We then use the additive Chernoff bound

$$\epsilon'_n(\delta/2, w) = \sqrt{\frac{\ln(2/\delta)}{2n}} \leq \sqrt{\frac{\ln(2N_{LB}(\epsilon, \mathcal{G}, L_1(\mathcal{D}))/\delta)}{2n}}$$

for an arbitrary $\epsilon > 0$.
The conditions of Lemma 11 hold.
We can then use the above upper bound on $\epsilon'_n(\delta/2, w)$ to simplify the result of Lemma 11, and take the minimum over ϵ to obtain the first desired bound of the corollary.

A Simple Example

We consider a one dimensional classification problem, where the input x is uniformly distributed in $[0,1]$, and the output $y \in \{\pm 1\}$ is generated according to

$$\Pr(y = 1|x) = \begin{cases} p & \text{if } x \geq w_* \\ (1-p) & \text{otherwise} \end{cases} \tag{11}$$

for some unknown $w_* \in [0,1]$ and $p \in (0.5, 1]$. See Figure 1.

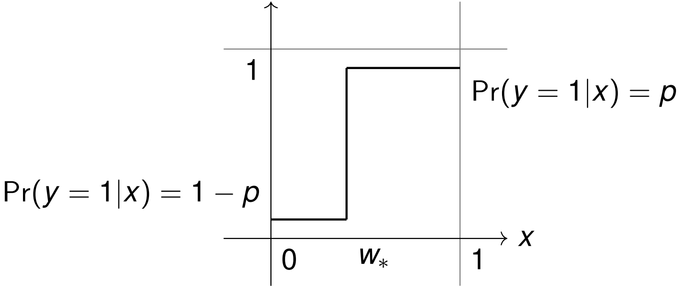

Figure: Conditional probability $\Pr(y = 1|x)$ as a function of x

A Simple Example (cont)

We don't know w_*, and consider a family of classifiers

$$f(w, x) = 2\mathbb{1}(x \geq w) - 1 = \begin{cases} 1 & \text{if } x \geq w \\ -1 & \text{otherwise} \end{cases},$$

where $w \in \Omega = [0, 1]$ is the model parameter to be learned from the training data. Here $\mathbb{1}(\cdot)$ is the binary indicator function.
In this example, we consider the following classification error loss

$$L(f(x), y) = \mathbb{1}(f(x) \neq y).$$

In this case, the optimal Bayes classifier is $f_*(x) = 2\mathbb{1}(x \geq w_*) - 1$, and the optimal Bayes error is

$$\mathbb{E}_{X,Y}\, L(f(w_*, X), Y) = 1 - p.$$

Lower Bracketing Cover

Given any $\epsilon > 0$, we let $w_j = 0 + j\epsilon$ for $j = 1, \ldots, \lceil 1/\epsilon \rceil$. Let

$$\phi_j(z) = \begin{cases} 0 & \text{if } x \in [w_j - \epsilon, w_j] \\ \phi(w_j, z) & \text{otherwise,} \end{cases}$$

where $z = (x, y)$. Note that $\phi_j \notin \mathcal{G}$.
It follows that for any $w \in [0, 1]$, if we let w_j be the smallest j such that $w_j \geq w$, then we have $\phi_j(z) = 0 \leq \phi(w, z)$ when $x \in [w_j - \epsilon, w_j]$, and $\phi_j(z) = \phi(w, z)$ otherwise, where $z = (x, y)$. Moreover,

$$\mathbb{E}_{Z \sim \mathcal{D}}[\phi_j(Z) - \phi(w, Z)] = \mathbb{E}_{X \in [w_j - \epsilon, w_j]}[0 - \phi(w, Z)] \geq -\epsilon.$$

We thus have

$$N_{LB}(\epsilon, \mathcal{G}, L_1(\mathcal{D})) \leq 1 + \epsilon^{-1}.$$

Oracle Inequality

We have (by picking $\epsilon = 2/n$):

$$\inf_{\epsilon>0}\left[\epsilon + \sqrt{\frac{2\ln(2N_{LB}(\epsilon,\mathcal{G},L_1(\mathcal{D}))/\delta)}{n}}\right] \leq \frac{2}{n} + \sqrt{\frac{2\ln((n+2)/\delta)}{n}}.$$

This implies the following additive oracle inequality from Corollary 14 with $\phi(w,z) = L(f(w,x),y)$.

Oracle Inequality

With probability at least $1-\delta$,

$$\mathbb{E}_{(X,Y)\sim\mathcal{D}}L(f(\hat{w},X),Y) \leq (1-p) + \frac{2}{n} + \sqrt{\frac{2\ln((n+2)/\delta)}{n}}.$$

Better Bounds with Variance Condition

In this section, we show that better bounds can be obtained with Bernstein's inequality under the following condition.

Definition 15 (Variance Condition)

Given a function class $\mathcal{G}$. We say it satisfies the variance condition if there exists $c_0, c_1 > 0$ such that for all $\phi(z) \in \mathcal{G}$:

$$\mathrm{Var}_{Z\sim\mathcal{D}}(\phi(Z)) \leq c_0^2 + c_1 \mathbb{E}_{Z\sim\mathcal{D}}\phi(Z), \tag{12}$$

where we require that $\mathbb{E}_{Z\sim\mathcal{D}}\phi(Z) \geq -c_0^2/c_1$ for all $\phi \in \mathcal{G}$.

In applications, the following modification of the variance condition is often more convenient to employ

$$\mathbb{E}_{Z\sim\mathcal{D}}[\phi(Z)^2] \leq c_0^2 + c_1 \mathbb{E}_{Z\sim\mathcal{D}}\phi(Z). \tag{13}$$

Example I

Example 16 (Bounded Function)

Let $\mathcal{G} = \{\phi(\cdot) : \forall z,\ \phi(z) \in [0, 1]\}$. Then $\mathcal{G}$ satisfies the variance condition (13) with $c_0 = 0$ and $c_1 = 1$.

Example II

Example 17 (Convex Least Squares)

Consider the least squares method $L(f(x), y) = (f(x) - y)^2$, with bounded response: $L(f(x), y) \leq M^2$ for some $M > 0$. Let $\mathcal{F}$ be a convex function class (that is, for any $f_1, f_2 \in \mathcal{F}$, and $\alpha \in (0, 1)$, $\alpha f_1 + (1 - \alpha) f_2 \in \mathcal{F}$), and define the optimal function in $\mathcal{F}$ as:

$$f_{\text{opt}} = \arg\min_{f \in \mathcal{F}} \mathbb{E}_{(x,y)\sim\mathcal{D}} L(f(x), y). \tag{14}$$

Let $z = (x, y)$, and

$$\mathcal{G} = \{\phi(\cdot) : \phi(z) = L(f(x), y) - L(f_{\text{opt}}(x), y), f(x) \in \mathcal{F}\}.$$

Then $\mathcal{G}$ satisfies the variance condition (13) with $c_0 = 0$, and $c_1 = 4M^2$.

Example III

Example 18 (Non-convex Least Squares)

More generally, if $\mathcal{F}$ is bounded nonconvex function class with $f(x) \in [0, M]$ for all $f \in \mathcal{F}$. If we assume that $y \in [0, M]$, then the variance condition may not hold with f_{opt} in (14). However, if we replace f_{opt} by $f_*(x) = \mathbb{E}[Y|X = x]$ in the definition of $\mathcal{G}$ as follows:

$$\mathcal{G} = \{\phi(\cdot) : \phi(z) = L(f(x), y) - L(f_*(x), y), f(x) \in \mathcal{F}\},$$

then all functions in $\mathcal{G}$ satisfy the variance condition (13) with $c_0 = 0$, and $c_1 = 2M^2$. Note that in general f_* may not belong to $\mathcal{F}$. However if the problem is well-specified (that is, $f_*(x) \in \mathcal{F}$), then the variance condition holds with $f_{\mathrm{opt}} = f_*$.

Uniform Convergence (Bernstein)

Theorem 19 (Simplification of Thm 3.21)

Assume condition (12) *is satisfied with* $c_0 = 0$*. Moreover, assume that the condition of Bernstein inequality is satisfied with* $b > 0$ *and* $V = \mathrm{Var}(\phi(Z))$*, and* $\mathbb{E}_{Z\sim\mathcal{D}}\phi(Z) \geq 0$*.*
Then $\forall\delta \in (0,1)$*, with probability at least* $1-\delta$*, the following inequality holds for all* $\gamma \in (0,1)$ *and* $w \in \Omega$*:*

$$(1-\gamma)\phi(w,\mathcal{D}) \leq \phi(w,\mathcal{S}_n) + \epsilon_n^\gamma(\delta,\mathcal{G},\mathcal{D}),$$

$$\epsilon_n^\gamma(\delta,\mathcal{G},\mathcal{D}) = \inf_{\epsilon\in[0,\epsilon_0]}\left[(1-\gamma)\epsilon + \frac{(3c_1+2\gamma b)\ln(N_{LB}(\epsilon,\mathcal{G},L_1(\mathcal{D}))/\delta)}{6\gamma n}\right].$$

Note that we obtain an $O(1/n)$ uniform convergence rate.

Theorem 3.21 also handles $c_0 \neq 0$.

Oracle Inequality (Bernstein)

Corollary 20 (Cor 3.22)

Let

$$w_* = \arg\min_{w \in \Omega} \mathbb{E}_{(X,Y)\sim\mathcal{D}} L(f(w,X),Y),$$

and assume that the conditions of Theorem 19 hold with

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

Then, with probability at least $1-\delta$*, the approximate ERM method* (6) *satisfies the following oracle inequality*

$$\mathbb{E}_{(X,Y)\sim\mathcal{D}} L(f(\hat{w},X),Y) \leq \mathbb{E}_{(X,Y)\sim\mathcal{D}} L(f(w_*,X),Y) + 2(\epsilon_n^{0.5}(\delta,\mathcal{G},\mathcal{D}) + \epsilon'),$$

where $\epsilon_n^{\gamma}(\delta,\mathcal{G},\mathcal{D})$ *is defined in Theorem 19.*

Simple Example Revisited

Consider Example on Slide 32, with the following modified $\phi(w, z)$:

$$\phi(w, z) = \mathbb{1}(f(w, x) \neq y) - \mathbb{1}(f(w_*, x) \neq y),$$

and the functions $\phi'_j(z) = \phi_j(z) - \mathbb{1}(f(w_*, x) \neq y)$ form an ϵ lower-bracketing cover, where $\phi_j(z)$ is defined on Slide 33 as

$$\phi_j(z) = \begin{cases} 0 & \text{if } x \in [w_j - \epsilon, w_j] \\ \phi(w_j, z) & \text{otherwise.} \end{cases}$$

A slight generalized Theorem 19 (Theorem 3.21 in the book with $c_0 \neq 0$) hold for $\epsilon \leq \epsilon_0$ with $c_0^2 = O(\epsilon_0)$, $c_1 = O(1)$, $b = 2$. We obtain

$$\epsilon_n^\gamma(\delta/2, \mathcal{G}, \mathcal{D}) = O\left(\frac{\ln(n/\delta)}{n}\right).$$

Simple Example Revisited: Oracle Inequality

Corollary 20 implies the following oracle inequality.

Oracle Inequality with Fast Convergence Rate

With probability at least $1 - \delta$:

$$\mathbb{E}_{(X,Y)\sim\mathcal{D}} \mathbb{1}(f(\hat{w}, X) \neq Y) \leq (1 - p) + O\left(\frac{\ln(n/\delta)}{n}\right).$$

Note also that $\mathbb{E}_{(X,Y)\sim\mathcal{D}}\, \mathbb{1}(f(w_*, X) \neq Y) = 1 - p$. This shows the ERM method has generalization error converging to the Bayes error at a fast rate of $O(\ln n/n)$.

Example: Parametric Model

In general, for bounded parametric function classes with d real-valued parameters (such as linear models $f(w, x) = w^\top x$ defined on a compact subset of $\mathbb{R}^d$), we expect the entropy (see Section 5.2 in the book) to behave as

Covering for Parametric Model

$$\ln N_{LB}(\epsilon, \mathcal{G}, L_1(\mathcal{D})) = O(d \ln(1/\epsilon)).$$

Assume that (13) holds with $c_0 = 0$ and $c_1 > 1$. Then it can be shown that the generalization bound in Corollary 20 implies

Oracle Inequality with Fast Rate

$$\mathbb{E}_{\mathcal{D}} L(f(\hat{w}, X), Y) \leq \mathbb{E}_{\mathcal{D}} L(f(w_*, X), Y) + O\left(\frac{\ln(n^d/\delta)}{n}\right).$$

General Bracketing Number

Definition 21 (Bracketing Number)

Let $\mathcal{G} = \{\phi(w, \cdot) : w \in \Omega\}$ be a real-valued function class, equipped with a pseudometric d. We say

$$\mathcal{G}(\epsilon) = \{[\phi_1^L(z), \phi_1^U(z)], \ldots, [\phi_N^L(z), \phi_N^U(z)]\}$$

is an ϵ-bracket of $\mathcal{G}$ under metric d if for all $w \in \Omega$, there exists $j = j(w)$ such that $\forall z$:

$$\phi_j^L(z) \le \phi(w, z) \le \phi_j^U(z), \qquad d(\phi_j^L, \phi_j^U) \le \epsilon.$$

The ϵ-bracketing number is the smallest cardinality $N_{[]}(\epsilon, \mathcal{G}, d)$ of such $\mathcal{G}(\epsilon)$. The quantity $\ln N_{[]}(\epsilon, \mathcal{G}, d)$ is called ϵ bracketing entropy.

L_p Bracketing

Given a distribution $\mathcal{D}$ and $p \geq 1$, we define L_p-seminorm in function space as

$$\|f - f'\|_{L_p(\mathcal{D})} = \left[\mathbb{E}_{Z\sim\mathcal{D}}|f(Z) - f'(Z)|^p\right]^{1/p}. \tag{15}$$

It induces a pseudometric, denoted as $d = L_p(\mathcal{D})$, and the corresponding bracketing number is $N_{[]}(\epsilon, \mathcal{G}, L_p(\mathcal{D}))$.

Proposition 22 (Prop 3.28)

We have for all $p \geq 1$:

$$N_{LB}(\epsilon, \mathcal{G}, L_1(\mathcal{D})) \leq N_{[]}(\epsilon, \mathcal{G}, L_1(\mathcal{D})) \leq N_{[]}(\epsilon, \mathcal{G}, L_p(\mathcal{D})).$$

It follows that Theorem 13 and Theorem 19 apply for all $N_{[]}(\epsilon, \mathcal{G}, L_p(\mathcal{D}))$ with $p \geq 1$.

Summary (Chapter 3)

- PAC Learning and Uniform Convergence
- Chernoff Bound + Union Bound, logarithmic dependency on class size N
- Additive Chernoff bound: $O(1/\sqrt{n})$ convergence
- Multiplicative Chernoff bound: $O(1/n)$ convergence (see book)
- Uniform Convergence and Oracle Inequality for ERM
- Bracketing Cover implies Uniform Convergence
- Additive Chernoff: $O(1/\sqrt{n})$ convergence rate
- Multiplicative Chernoff: $O(1/n)$ convergence for realizable case (see book)
- Variance condition implies faster rate
- Bernstein: can lead to $O(1/n)$ rate for non-realizable cases