1 Introduction

1.1 Motivation

In the last lecture we consider a learning problem in which the optimal function belonged to a finite class of functions. Specifically, for some collection of functions \( \mathcal{F} \) with finite cardinality \( |\mathcal{F}| \leq \infty \), we have

\[
\min_{f \in \mathcal{F}} R(f) = 0 \Rightarrow f^* \in \mathcal{F}
\]

This is almost always not the situation in the real-world learning problems. Let us suppose we have a finite collection of candidate functions \( \mathcal{F} \). Furthermore, we do not assume that the optimal function \( f^* \), which satisfies

\[
R(f^*) = \inf_{f} R(f)
\]

, where the inf is taken over all measurable functions, is a member of \( \mathcal{F} \). That is, we make few, if any, assumptions about \( f^* \). This situation is sometimes termed as Agnostic Learning. The root of the word agnostic literally means not known. The term agnostic learning is used to emphasize the fact that often, perhaps usually, we may have no prior knowledge about \( f^* \). The question then arises about how we can reasonably select an \( f \in \mathcal{F} \) in this setting.

1.2 The Problem

The PAC style bounds discussed in the previous lecture, offer some help. Since we are selecting a function based on the empirical risk, the question is how close is \( \hat{R}_n(f) \) to \( R(f) \) \( \forall f \in \mathcal{F} \). In other words, we wish that the empirical risk is a good indicator of the true risk for every function in \( \mathcal{F} \). If this is case, the selection of \( f \) that minimizes the empirical risk

\[
\hat{f}_n = \arg \min_{f \in \mathcal{F}} \hat{R}_n(f)
\]

should also yield a small true risk, that is, \( R(\hat{f}_n) \) should be close to \( \min_{f \in \mathcal{F}} R(f) \). Finally, we can thus state our desired situation as

\[
P(|\hat{R}_n(f) - R(f)| > \epsilon) < \delta, \ \forall f \in \mathcal{F}
\]

In other words, \( \forall f \in \mathcal{F} \), with probability at least \( 1 - \delta \), \( |\hat{R}_n(f) - R(f)| > \epsilon \). In this lecture, we will start to develop bounds of this form. First we will focus on bounding \( P(|\hat{R}_n(f) - R(f)| > \epsilon) \) for one fixed \( f \in \mathcal{F} \).
2 Developing Initial Bounds

To begin, let us recall the definition of empirical risk for \( \{X_i, Y_i\}_{i=1}^n \) be a collection of training data. Then the empirical risk is defined as

\[
\hat{R}_n(f) = \frac{1}{n} \sum_{i=1}^{n} \ell(f(X_i), Y_i)
\]

Note that since the training data \( \{X_i, Y_i\}_{i=1}^n \) are assumed to be i.i.d. pairs, each term in the sum is an i.i.d random variables.

Let

\[
L_i = \ell(f(X_i), Y_i)
\]

The collection of losses \( \{L_i\}_{i=1}^n \) is i.i.d according to some unknown distribution (depending on the unknown joint distribution of (X, Y) and the loss function). The expectation of \( L_i \) is

\[
E[\ell(f(X_i), Y_i)] = E[\ell(f(X), Y)] = R(f),
\]

the true risk of \( f \). For now, let’s assume that \( f \) is fixed.

\[
E[\hat{R}_n(f)] = \frac{1}{n} \sum_{i=1}^{n} E[\ell(f(X_i), Y_i)] = \frac{1}{n} \sum_{i=1}^{n} E[L_i] = R(f)
\]

We know from the strong law of large numbers that the average (or empirical mean) \( \hat{R}_n(f) \) converges almost surely to the true mean \( R(f) \). That is, \( \hat{R}_n(f) \to R(f) \) almost surely as \( n \to \infty \). The question is how fast.

3 Concentration of Measure Inequalities

Concentration inequalities are upper bounds on how fast empirical means converge to their ensemble counterparts, in probability.

Area of the shaded tail regions is \( P(|\hat{R}_n(f) - R(f)| > \epsilon) \). We are interested in finding out how fast this probability tends to zero as \( n \to \infty \).

At this stage, we recall Markov’s Inequality. Let \( Z \) be a nonnegative random variable.

\[
E[Z] = \int_0^\infty z p(z)dz = \int_0^t z p(z)dz + \int_t^\infty z p(z)dz \geq 0 + t \int_t^\infty z p(z)dz = t P(Z \geq t)
\]

\[
\Rightarrow P(Z \geq t) \leq \frac{E[Z]}{t}
\]

\[
\Rightarrow P(Z^2 \geq t^2) \leq \frac{E[Z^2]}{t^2}
\]

Take

\[
Z = |R_n(f) - R(f)| \quad \text{and} \quad t = \epsilon
\]
Figure 1: Distribution of $\hat{R}_n(f)$

\[
P(\{|\hat{R}_n(f) - R(f)| \geq \epsilon\}) \leq \frac{E[|\hat{R}_n(f) - R(f)|^2]}{\epsilon^2} \leq \frac{\text{var}(\hat{R}_n(f))}{\epsilon^2} = \frac{\sum_{i=1}^{n} \text{var}(L_i)}{\epsilon^2} = \frac{\text{var}(\ell(X,Y))}{n\epsilon^2} = \frac{\sigma_L^2}{n\epsilon^2}
\]

So, the probability goes to zero at a rate of at least $n^{-1}$. However, it turns out that this is an extremely loose bound. According to the Central Limit Theorem

\[
\hat{R}_n(f) = \frac{1}{n} \sum_{i=1}^{n} L_i \rightarrow N\left(R(f), \frac{\sigma_L^2}{n}\right) \quad \text{as} \quad n \rightarrow \infty
\]

in distribution. This suggests that for large values of $n$,

\[
P(\{|\hat{R}_n(f) - R(f)| \geq \epsilon\}) \approx O\left(e^{-\frac{\sigma_L^2}{2\epsilon^2}}\right)
\]

That is, the Gaussian tail probability is tending to zero exponentially fast.
4 A Dichotomy

Obviously, the bound based on Markov’s inequality is extremely loose for large. Tighter concentration inequalities can be derived using more sophisticated techniques. There is an important dichotomy at this point into the class of bounded loss functions (leading to bounded random variables $L_i$) and unbounded loss functions (leading to unbounded random variables $L_i$).

Example 1 Bounded Loss Functions

By this, we mean any loss function mapping into a bounded set, for example, 
$$
\ell : \mathcal{Y} \times \mathcal{Y} \to [0, 1]
$$

0–1 loss, $R(f) = E[1_{f(X) \neq Y}] = P(f(X) \neq Y)$.

So here, $L_i = 0$ or $1$.

Example 2 Unbounded Loss Functions

Any loss function mapping into an unbounded set, for example squared error, $R(f) = E[(f(X) - Y)^2]$.

The case of unbounded losses is simpler, since we can exploit the boundedness in a key way. Therefore, we can concentrate on bounded loss functions and classification problems first, and later we will look at unbounded losses and estimation problems.

5 Bounded Loss Functions and Chernoff’s Bound

Note that for any nonnegative random variable $Z$ and $t > 0$,

$$
P(Z \geq t) = P(e^{sZ} \geq e^{st}) \leq \frac{E[e^{sZ}]}{e^{st}}, \forall s > 0 \text{ by Markov’s inequality}
$$

Chernoff’s bound is based on finding the value of $s$ that minimizes the upper bound. If $Z$ is a sum of independent random variables. For example, say

$$
Z = \sum_{i=1}^{n} (\ell(f(X_i), Y_i) - R(f)) = n \left( \hat{R}_n(f) - R(f) \right)
$$

then the bound becomes

$$
P \left( \sum_{i=1}^{n} (L_i - E[L_i]) \geq t \right) \leq e^{-st} E[e^{s \sum_{i=1}^{n} (L_i - E[L_i])}] \leq e^{-st} \prod_{i=1}^{n} E[e^{s(L_i - E[L_i])}], \text{ from independence.}
$$

Thus, the problem of finding a tight bound boils down to finding a good bound for $E[s^{s(L_i - E[L_i])}]$. Chernoff (’52), first studied this situation for binary random variables. Then, Hoeffding (’63) derived a more general result for arbitrary bounded random variables.

Theorem 1 Hoeffding’s Inequality

Let $Z_1, Z_2, ..., Z_n$ be independent bounded random variables such that $Z_i \in [a_i, b_i]$ with probability 1. Let $S_n = \sum_{i=1}^{n} Z_i$. Then for any $t > 0$, we have

$$
P(\|S_n - E[S_n]\| \geq t) \leq 2e^{-\frac{t^2}{\sum_{i=1}^{n} (b_i - a_i)^2}}$$
Application: Let $Z_i = 1_{f(X_i) \neq Y_i} - R(f)$, as in the classification problem. Then for a fixed $f$, it follows from Hoeffding’s inequality (i.e., Chernoff’s bound in this special case) that

$$P(\hat{R}_n(f) - R(f) \geq \epsilon) = P\left(\frac{1}{n}|S_n - E[S_n]| \geq \epsilon\right)$$

$$= P(|S_n - E[S_n]| \geq n\epsilon)$$

$$\leq 2e^{-\frac{2n\epsilon^2}{n}}$$

$$= 2e^{-2n\epsilon^2}$$

Proof: The key to proving Hoeffding’s inequality is the following upper bound: if $Z$ is a random variable with $E[Z] = 0$ and $a \leq Z \leq b$, then

$$E[e^{sZ}] \leq e^{\frac{2(a-s)^2}{8}}$$

This upper bound is derived as follows. By the convexity of the exponential function,

$$e^{sz} \leq \frac{z-a}{b-a} e^{s(b-a)} + \frac{b-z}{b-a} e^{sa}$$

for $a \leq z \leq b$

Figure 2: Convexity of exponential function.

Thus,

$$E[e^{sZ}] \leq E\left[\frac{Z-a}{b-a} e^{s(b-a)} + E\left[\frac{b-Z}{b-a}\right] e^{sa}\right]$$

$$= \frac{b}{b-a} e^{sa} - \frac{a}{b-a} e^{sb}$$

since $E[Z] = 0$

$$= (1 - \theta + \theta e^{s(b-a)}) e^{-\theta s(b-a)}$$

where $\theta = \frac{-a}{b-a}$
Now let \( u = s(b - a) \) and define \( \phi(u) \equiv -\theta u + \log(1 - \theta + \theta e^u) \)

Then we have

\[
E[e^{sZ}] \leq (1 - \theta + \theta e^{s(b-a)})e^{-\theta s(b-a)} = e^{\phi(u)}
\]

To minimize the upper bound let’s express \( \phi(u) \) in a Taylor’s series with remainder:

\[
\phi(u) = \phi(0) + u\phi'(0) + \frac{u^2}{2} \phi''(v) \text{ for some } v \in [0, u]
\]

\[
\phi'(u) = -\theta + \frac{\theta e^u}{1 - \theta + \theta e^u} \Rightarrow \phi'(u) = 0
\]

\[
\phi''(u) = \frac{\theta e^u}{1 - \theta + \theta e^u} - \frac{\theta e^u}{(1 - \theta + \theta e^u)^2}
= \frac{\theta e^u}{1 - \theta + \theta e^u}(1 - \frac{\theta e^u}{1 - \theta + \theta e^u})
= \rho(1 - \rho)
\]

Now, \( \phi''(u) \) is maximized by

\[
\rho = \frac{\theta e^u}{1 - \theta + \theta e^u} = \frac{1}{2} \Rightarrow \phi''(u) \leq \frac{1}{4}
\]

So,

\[
\phi(u) \leq \frac{u^2}{8} = \frac{s^2(b - a)^2}{8}
\]

\[
\Rightarrow E[e^{sZ}] \leq e^{\frac{s^2(b - a)^2}{8}}
\]

Now, we can apply this upper bound to derive Hoeffding’s inequality.

\[
P(S_n - E[S_n] \geq t) \leq e^{-st} \prod_{i=1}^{n} E[e^{s(L_i - E[L_i])}]
= e^{-st} \prod_{i=1}^{n} e^{\frac{s^2(b_i - a_i)^2}{2}}
= e^{-st} e^{\frac{s^2}{8} \sum_{i=1}^{n} (b_i - a_i)^2}
= e^{\frac{s^2}{8} \sum_{i=1}^{n} (b_i - a_i)^2}
\text{by choosing } s = \frac{4t}{\sum_{i=1}^{n} (b_i - a_i)^2}
\]

Similarly, \( P(E[S_n] - S_n \geq t) \leq e^{\frac{-t^2}{8 \sum_{i=1}^{n} (b_i - a_i)^2}} \). This completes the proof of the Hoeffding’s theorem.
Now, we want a bound like this to hold for all \( f \in \mathcal{F} \). Let us enumerate the functions in \( \mathcal{F} \) as \( f_1, f_2, \ldots, f_{|\mathcal{F}|} \), where \( |\mathcal{F}| \) denotes the cardinality of \( \mathcal{F} \). We would like to bound the probability that \( |\hat{R}_n(f) - R(f)| \geq \epsilon \) for any \( f \in \mathcal{F} \). This probability is

\[
P\left( |\hat{R}_n(f_1) - R(f_1)| \geq \epsilon \text{ or } \cdots \text{ or } |\hat{R}_n(f_{|\mathcal{F}|}) - R(f_{|\mathcal{F}|})| \geq \epsilon \right) = P\left( \bigcup_{f \in \mathcal{F}} |\hat{R}_n(f) - R(f)| \geq \epsilon \right).
\]

\[
P\left( \bigcup_{f \in \mathcal{F}} |\hat{R}_n(f) - R(f)| \geq \epsilon \right) \leq \sum_{f \in \mathcal{F}} P(|\hat{R}_n(f) - R(f)| \geq \epsilon), \text{ the "union of events" bound}
\]

\[
\leq 2|\mathcal{F}|e^{-2n\epsilon^2}, \text{ by Hoeffding’s inequality.}
\]

Thus, we have shown that \( \forall f \in \mathcal{F} \) with probability at least \( 1 - 2|\mathcal{F}|e^{-2n\epsilon^2} \),

\[
|\hat{R}_n(f) - R(f)| < \epsilon.
\]

And accordingly, we can be reasonably confident in selecting \( f \) from \( \mathcal{F} \) based on the empirical risk function \( \hat{R}_n \).