8.1 Uniform Laws

\begin{equation}
L(\theta, \theta^*) = E_{x \sim p(\cdot | \theta^*)} \ell(x, \theta)
\end{equation}

\begin{equation}
L_n(\theta, \theta^*) = \frac{1}{n} \sum_{i=1}^{n} \ell(x_i, \theta)
\end{equation}

Empirical Risk Minimization (ERM) is what we actually minimize using samples

\[ \hat{\theta} \in \arg \inf_{\theta \in \Theta_0 \subseteq \Theta} L_n(\theta, \theta^*) \]

Our “gold standard” is the optimum w.r.t. the true expectation:

\[ \theta_0 \in \arg \inf_{\theta \in \Theta_0 \subseteq \Theta} L(\theta, \theta^*) \]

To compare these two quantities, we will look at the excess:

\[ E(\hat{\theta}, \theta_0) = L(\hat{\theta}, \theta^*) - L(\theta_0, \theta^*) \]

\begin{equation}
= L(\hat{\theta}, \theta^*) - L_n(\hat{\theta}, \theta^*) + L_n(\hat{\theta}, \theta^*) - L_n(\theta_0, \theta^*) + L_n(\theta_0, \theta^*) - L(\theta_0, \theta^*) \tag{8.6}
\end{equation}

We know \( T_2 \leq 0 \) by definition of \( \hat{\theta} \) being optimal for \( L_n \).

We can bound \( T_3 \) directly using a tail bound:

\begin{equation}
L_n(\theta_0, \theta^*) - L(\theta_0, \theta^*) = \frac{1}{n} \sum_{i=1}^{n} \ell(x_i, \theta_0) - E[\ell(x_i, \theta_0)] \tag{8.7}
\end{equation}

For \( T_3 \) we assumed the \( x_i \) were iid, so each \( \ell(x_i, \theta) \) was independent. For \( T_1, \hat{\theta} \) depends on \( x_i \), each \( \ell(x_i, \hat{\theta}) \) is dependent. Thus, we cannot directly apply the tail bounds we derived in the last lecture.

\begin{equation}
L_n(\hat{\theta}, \theta^*) - L(\hat{\theta}, \theta^*) = \frac{1}{n} \sum_{i=1}^{n} \ell(x_i, \hat{\theta}) - E_{x \sim p(\cdot | \theta^*)}[\ell(x, \hat{\theta})] \tag{8.8}
\end{equation}

\begin{equation}
\leq \sup_{\theta \in \Theta_0} \left| \sum_{i=1}^{n} \ell(x_i, \theta) - E[\ell(x_i, \theta)] \right| \triangleq \delta_n \tag{8.9}
\end{equation}

Noting that we can also bound \( T_3 \leq \delta_n \), we obtain the following bound on the excess:

\[ E(\hat{\theta}, \theta_0) \leq 2\delta_n \tag{8.10} \]
8.1.1 Uniform Laws

We’ll begin by defining \( x^\theta \) as a random variable drawn from \( p^\theta (\cdot) \). We are interested in the deviation between the sample mean \( \frac{1}{n} \sum_{i=1}^{n} x_i^\theta \) and its expectation, \( \mathbb{E}[x^\theta] \). In particular, we are interested in the maximum deviation between these quantities, as we vary \( \theta \):

\[
\sup_{\theta} \left| \frac{1}{n} \sum_{i=1}^{n} x_i^\theta - \mathbb{E}[x^\theta] \right|
\]

(8.11)

8.1.2 Uniform Laws for CDFs

One early application of uniform laws was to cumulative density functions (CDFs):

\[
F(t) \triangleq \mathbb{P}(x \leq t) = \mathbb{E}[\mathbb{1}(x \in (-\infty, t))]
\]

(8.12)

We now define the empirical CDF as

\[
F_n(t) \triangleq \frac{1}{n} \sum_{i=1}^{n} \mathbb{1}(x_i \leq t)
\]

(8.13)

For a fixed \( t \), the Law of Large Numbers tells us that the empirical CDF converges to the true CDF as \( n \) goes to infinity:

\[
F_n(t) \xrightarrow{a.s.} F(t)
\]

(8.14)

But we are really interested in the CDF converges simultaneously for all \( t \).

Theorem 8.1 (Glivenko-Cantelli)  This theorem tells us that CDFs converge uniformly

\[
\|F_n - F\|_\infty \xrightarrow{a.s.} 0
\]

(8.15)

where \( \|F - G\|_\infty \triangleq \sup_{t \in \mathbb{R}} \|F(t) - G(t)\| \)

However, the Glivenko-Cantelli theorem does not tell us about uniform convergence of other quantities. Now, we will prove a generalization of the Glivenko-Cantelli theorem (that will include the result we want to ERM). We consider iid samples \( x_i \sim \mathbb{P} \), where each sample belongs to some set: \( x_i \in \mathcal{X} \). We consider a set of functions \( \mathcal{F} \) defined over the set \( \mathcal{X} \). We are interested in the following deviation:

\[
\sup_{f \in \mathcal{F}} \left| \frac{1}{n} \sum_{i=1}^{n} f(x_i) - \mathbb{E}[f(x)] \right|
\]

(8.16)

The Glivenko-Cantelli theorem was a special case, where we considered the following set of functions:

\[
\mathcal{F} = \{ \mathbb{1}(x \in (-\infty, t)) : t \in \mathbb{R} \}
\]

(8.17)

For ERM, we consider another set of functions:

\[
\mathcal{F} = \{ \ell(\cdot, \theta) : \theta \in \Theta \}
\]

(8.18)

Definition: We define the distance \( \| : \|_{\mathcal{F}} \) as the maximum absolute value over functions in \( \mathcal{F} \):

\[
\|\mathbb{P}_n - \mathbb{P}\|_{\mathcal{F}} = \sup_{f \in \mathcal{F}} \left| \frac{1}{n} \sum_{i=1}^{n} f(x_i) - \mathbb{E}[f(x)] \right|
\]

(8.19)
Definition: A set of functions $\mathcal{F}$ is a Glivenko-Cantelli Class if the following result holds for all distributions $P$:

$$
\|P_n - P\|_{\mathcal{F}} \xrightarrow{\text{prob.}} 0
$$

We say that $\mathcal{F}$ is a strong Glivenko-Cantelli Class if we have almost-sure convergence:

$$
\|P_n - P\|_{\mathcal{F}} \xrightarrow{a.s.} 0
$$

Example: The set $\mathcal{F} = \{ 1(x \in S) ; S \subseteq [0, 1] \}$ is not a Glivenko-Cantelli Class. If we draw samples $x$ from a continuous density.

$$
\sup_{S \subseteq [0, 1]} |E_n[1(x \in S)] - E[1(x \in S)]| = 1 \neq 0
$$

Next, we will look at determining whether a function class is Glivenko-Cantelli. To do this, we will only look at the function evaluations, rather than the functions themselves:

$$
\mathcal{F}(x_1^n) \triangleq \{ (f(x_1), f(x_2), \ldots, f(x_n)) ; f \in \mathcal{F} \} \subseteq \mathbb{R}^n
$$

Intuitively, if the function only takes a few values, that it is more likely that the maximum deviation between the expected value and the empirical average will be small. We recall the definition of the Rademacher Complexity:

$$
R(S) \triangleq E_\epsilon \left[ \sup_{a \in S} \left| \frac{1}{n} \sum_{i=1}^{n} \epsilon_i a_i \right| \right]
$$

If the set $S$ is small, then it is unlikely that we can find a vector $a \in S$ that has high correlation with the noise vector $\epsilon$. As we increase the size of $S$, we expect that it will be more likely to find a vector in $S$ with high correlation.

We now will look at the Rademacher complexity of the set of function evaluations. The empirical Rademacher complexity is

$$
R\left( \frac{\mathcal{F}(x_1^n)}{n} \right) = E_\epsilon \left[ \sup_{f \in \mathcal{F}} \left| \frac{1}{n} \sum_{i=1}^{n} \epsilon_i f(x_i) \right| \right]
$$

We can also look at the population Rademacher complexity by taking an expectation over the samples. This quantity is also called the Rademacher complexity of the function class $\mathcal{F}$:

$$
R_n(\mathcal{F}) = E_{x^n} \left[ R\left( \frac{\mathcal{F}(x_1^n)}{n} \right) \right]
$$

Theorem 8.2 Let a function class $\mathcal{F}$ that is $b$-uniformly bounded (i.e. $\|f\|_{\infty} \leq b, \forall f \in \mathcal{F}$) be given. Then, for all $n \geq 1, \delta \geq 0$, we have

$$
\|P_n - P\|_{\mathcal{F}} \leq 2R_n(\mathcal{F}) + \delta
$$

with probability at least $1 - \exp\left(\frac{-n\delta^2}{2b^2}\right)$.

An immediately corollary of this theorem is that if the Rademacher complexity $R_n(\mathcal{F})$ converges to zero, then the function class $\mathcal{F}$ is a GC class.

Theorem 8.3 Let a $b$-uniformly bounded function class $\mathcal{F}$ be given. Then, for all $n \geq 1, \delta \geq 0$, we have

$$
\|P_n - P\|_{\mathcal{F}} \geq \frac{1}{2} R_n(\mathcal{F}) - \sup_{f \in \mathcal{F}} |E[f]| - \delta
$$

with probability at least $1 - \exp\left(\frac{-n\delta^2}{2b^2}\right)$.

Taken together, these results say that the Rademacher complexity gives us both upper and lower bounds for the maximum deviation. Thus, we need to find a way to bound the Rademacher complexity.
8.2 Polynomial Discrimination

**Definition:** A function class \( \mathcal{F} \) has *polynomial discrimination* on the order \( v \geq 1 \) if, for all \( x^n \in \mathcal{X}^n \), we have the following bound on the cardinality of function evaluations:

\[
\text{card}(\mathcal{F}(x^n_i)) \leq (n + 1)^v
\]  

(8.29)

Note that \( \mathcal{F}(x^n_i) \) is a set containing length-\( n \) vectors. We are counting the number of unique vectors in this set. For example, if each function \( f \in \mathcal{F} \) is binary, then there are at most \( 2^n \) bit vectors, so

\[
\text{card}(\mathcal{F}(x^n_i)) \leq 2^n
\]  

(8.30)

Noting that \( 2^n \) is exponential in \( n \), not polynomial, we see that arbitrary binary functions are not polynomial discriminable.

**Theorem 8.4** Let a function class \( \mathcal{F} \) that is polynomial discriminable with order \( v \) be given. Then we can bound the Rademacher complexity of \( \mathcal{F} \) as follows:

\[
R_n(\mathcal{F}) \leq 2 \left( E_{x^n_1}[D(x^n_1)] \right) \sqrt{\frac{v \log(n + 1)}{n}} \quad \text{where} \quad D(x^n_1) \triangleq \sup_{f \in \mathcal{F}} \sqrt{\frac{1}{n} \sum_{i=1}^{n} f^2(x_i)}
\]  

(8.31)

**Example:** Let’s look at the class of CDFs, \( \mathcal{F} = \{ \mathbb{1}(x \in (-\infty, t]) : t \in \mathbb{R} \} \). Let’s further assume that our samples \( x^n_i \) are sorted:

\[
x_1 \leq x_2 \leq \cdots \leq x_n
\]  

(8.32)

For a fixed \( t \), we know that \( \mathbb{1}(x \in (-\infty, t]) \) will be 1 for small \( i \) and 0 for large \( i \). Thus, there are \( n + 1 \) possible values for the vector \( \mathbb{1}(x^n_i \in (-\infty, t]) \).