In this lecture, the focus is on bounding the generalization error when $H$ is infinite. We have the following theorems

**Theorem 1.** Given a function class $H$ (including infinite $H$), $\forall \epsilon, \delta > 0, \forall D, \forall h^* \in H$, given sample $S: (x_1, y_1), \ldots, (x_m, y_m) \sim D$, let $h_{erm}$ be the output of ERM. Then with probability $\geq 1 - \delta$, $\text{err}(h_{erm}) \leq \epsilon$, provided $m \geq \frac{2}{\epsilon^2} \log\left(\frac{2C[2m]}{\delta}\right)$.

**Theorem 2.** Given $H$ (including infinite $H$), $\forall \epsilon, \delta > 0, \forall D, h^*$, given $S$, with probability $\geq 1 - \delta, \forall h \in H$, $|\text{err}_s(h) - \text{err}(h)| \leq \epsilon$, if $m \geq \frac{8}{\epsilon^2} \log\left(\frac{2C[2m]}{\delta}\right)$.

We will prove Theorem 2 and in the process introduce a general technique called symmetrization to bound the over-fitting error. Before that, we will need the following basic tail inequalities.

**Theorem 3** (Hoeffding’s Inequality). Given independent random variables $X_1, \ldots, X_m \in [0, 1]$, where each r.v. has the same mean: $E(X_i) = \mu$, $\Pr(|\frac{1}{m}\sum_{i=1}^{m} X_i - \mu| \geq \epsilon) \leq e^{-2m\epsilon^2}$, $\forall \epsilon > 0$.

We will also need a generalization of the above called McDiarmid’s inequality that bounds deviations of functions of random variables from their expected value.

**Theorem 4** (McDiarmid’s inequality). Let $X_1, \ldots, X_m$, be independent random variables taking values in the set $X$. Let $f(X_1, X_2, \ldots, X_m): X^m \rightarrow R$ be a $c$-lipschitz function. Then $\Pr(\|f(X_1, \ldots, X_m) - E[f(X_1, \ldots, X_m)]\| > \epsilon) \leq e^{\frac{-2\epsilon^2}{c^2}}$, $\forall \epsilon > 0$.

Note that a function is $c$-lipschitz if its value changes at most $c$ when any single variable changes within its domain. In other words, $|f(X_1, \ldots, X_i, \ldots, X_m) - f(X_1, \ldots, X_i', \ldots, X_m')| \leq c, \forall i$ and $\forall X_i, X_i' \in X$. Notice that by restricting $X_i$ to be in $\{0, 1\}$, and taking $f(X_1, \ldots, X_m) = m \sum_{i=1}^{m} X_i$ and $c = \frac{1}{m}$, we can recover Hoeffding’s inequality. Now onto the proof of Theorem 2. We will prove the following restatement of the theorem: with probability $\geq 1 - \delta$, $\sup_{h \in H} |\text{err}_s(h) - \text{err}(h)| \leq 2 \sqrt{\frac{\log C[2m]}{m}} + \sqrt{\frac{\log(\frac{1}{\delta})}{2m}}$.

**Proof of Theorem 3** Let $err_s(h) = \frac{1}{m} \sum_{i=1}^{m} I(h(x_i) \neq y_i) = \frac{1}{m} \sum_{i=1}^{m} g(h(x_i, y_i))$. Then, $err(h) = E_{X \sim D}[g(h(x, y))]$. The proof of the claim can be further decomposed into two steps. The first step is to show $E_S[\sup_{h \in H} |\sum_{i=1}^{m} \frac{g(h(x_i, y_i))}{m} - E[g(h(x, y))]|] \leq 2 \sqrt{\frac{\log C[2m]}{m}}$, the second step is to show that with probability $\geq 1 - \delta$, $\sup_{h \in H} |\sum_{i=1}^{m} \frac{g(h(x_i, y_i))}{m} - E[g(h(x, y))]| = \sup_{h \in H} |\text{err}_s(h) - \text{err}(h)| \leq 2 \sqrt{\frac{\log C[2m]}{m}} + \sqrt{\frac{\log(\frac{1}{\delta})}{2m}}$. Suppose the claim of step 1 holds, then by applying McDiarmid’s inequality, the claim of step 2 also holds: let $F(x_1, y_1, \ldots, x_m, y_m) = \sup_{h \in H} |\sum_{i=1}^{m} \frac{g(h(x_i, y_i))}{m} - E[g(h(x, y))]|$. It is evident that $F$ is $c$-lipschitz where $c = \frac{1}{m}$. Applying McDiarmid’s inequality by setting $\epsilon = \sqrt{\frac{\log(\frac{1}{\delta})}{2m}}$, one has $\Pr(||F(x_1, y_1, \ldots, x_m, y_m) - E[F(x_1, y_1, \ldots, x_m, y_m)]|| \geq \sqrt{\frac{\log(\frac{1}{\delta})}{2m}}) \leq \exp(-\frac{2\log(\frac{1}{\delta})m^2}{2m^2}) = e^{-\log(\frac{1}{\delta})} = \delta \Rightarrow \text{w.p.} \geq 1 - \delta, ||F(x_1, y_1, \ldots, x_m, y_m) - E[F(x_1, y_1, \ldots, x_m, y_m)]|| \leq \sqrt{\frac{\log(\frac{1}{\delta})}{2m}}$. Therefore, the problem is to
show the claim of step 1. To this end, two tricks are used. One is to approximate
the expectation by another random sample. The other is to change expectation by introducing
random signs. Recall again that we need to show:\footnote{We have removed the absolute value here. The other direction can be proved similarly.}

\begin{equation}
E_S[\sup_{h \in H}(E[gh(x, y)] - \sum_{i=1}^{m} \frac{g_h(x_i, y_i)}{m})] \leq 2\sqrt{\frac{\log C[2m]}{m}} \tag{1}
\end{equation}

**Symmetrization:** Approximate $E_{X \sim D}[g_h(x, y)]$ by another random sample $S' = \{(x'_1, y'_1), \ldots, (x'_m, y'_m)\}$. Since $E_{X \sim D}[g_h(x, y)] = \frac{1}{m} \sum_{i=1}^{m} E_{S'}[g_h(x'_i, y'_i)]$, we get that

\begin{equation}
[1] = E_S[\sup_{h \in H} (\frac{1}{m} \sum_{i=1}^{m} E_{S'}[g_h(x'_i, y'_i)] - \frac{1}{m} \sum_{i=1}^{m} g_h(x_i, y_i))] 
= E_S[\sup_{h \in H} E_{S'}[\frac{1}{m} \sum_{i=1}^{m} (g_h(x'_i, y'_i) - g_h(x_i, y_i))] 
\leq E_S E_{S'}[\sup_{h \in H} (\frac{1}{m} \sum_{i=1}^{m} (g_h(x'_i, y'_i) - g_h(x_i, y_i)))] \tag{2}
\end{equation}

**Rademacher Averages:** Imagine the following equivalent process for generating samples $S$ and $S'$: Draw a set $T$ of $2m$ random examples. For each example $x_i \in T$, draw a random $\pm \sigma_i$ to decide whether the example falls in $S$ or $S'$. Here $\sigma_i$ is known as a Rademacher random variable and takes values in $\{+1, -1\}$ with equal probability. It follows that

\begin{equation}
E_S E_{S'}[\sup_{h \in H} (\frac{1}{m} \sum_{i=1}^{m} (g_h(x'_i, y'_i) - g_h(x_i, y_i))] = E_{T=S\cup S'} E_{\sigma}[\sup_{h \in H} (\frac{1}{m} \sum_{i=1}^{2m} \sigma_i g_h(x_i, y_i))] 
\leq E_T E_{\sigma}[\sup_{h \in H} (\frac{1}{m} \sum_{i=1}^{2m} \sigma_i (h(x_i) + y_i))] 
= 2E_T E_{\sigma}[\sup_{h \in H} (\frac{1}{2m} \sum_{i=1}^{2m} \sigma_i h(x_i))] 
\triangleq 2R_{2m}(H) \tag{3}
\end{equation}

In the second inequality we have used the fact that $g_h(x_i, y_i) \leq h(x_i) + y_i$. The quantity $R_{2m}(H)$ is known as the Rademacher complexity of the function class $H$. Note that it is distribution dependent quantity, unlike the VC dimension. The higher $R_{2m}(H)$ is, the higher the chance that a function in $H$ overfits. In addition, the claim $R_{2m}(H) \leq \sqrt{\frac{\log C[2m]}{m}}$ immediately implies that equation $[1]$ holds. This claim is proved in Lemma $[1]$. \hfill \square

**Lemma 1** (Massart’s Finite Class Lemma). Let $A \subset \mathbb{R}^m$ be a finite set, let $R = \sup_{a \in A}(\sum_{i=1}^{m} a_i^2)^{1/2}$. Then,

\begin{equation}
E_{\sigma}[\sup_{a \in A} (\frac{1}{m} \sum_{i=1}^{m} \sigma_i a_i)] \leq \frac{R \sqrt{2 \log |A|}}{m}
\end{equation}

We can use Massart’s lemma to bound $R_{2m}(H)$ as follows. Notice that $R_{2m}(H)$, $R_{2m}(H) = E_T E_{\sigma}[\sup_{h \in H} (\frac{1}{2m} \sum_{i=1}^{2m} \sigma_i h(x_i))] = E_T E_{\sigma}[\sup_{h \in H[2m]} (\frac{1}{2m} \sum_{i=1}^{2m} \sigma_i h(x_i))] [T]$. \hfill \square
Here $H[2m]$ denotes the finite set of functions that result in at most $C[2m]$ different labels on $T$ using $H$. Now apply Massart’s lemma with $A = \{(h(x_1), \ldots, h(x_{2m})) | h \in H[2m]\}$. It is easy to see that $R = \sup_{a \in A} (\sum_{i=1}^{2m} a_i^2)^{1/2} = \sqrt{2m}$.

**Proof of Lemma 1.** For any $s > 0$, using Jensen’s inequality we get that

$$\exp(s \sup_{a \in A} (\sum_{i=1}^{m} \sigma_i a_i)) \leq \sup_{a \in A} (\exp(s \sum_{i=1}^{m} \sigma_i a_i))$$

(monotonicity of functions) $\leq \sum_{a \in A} (\exp(s \sum_{i=1}^{m} \sigma_i a_i))$

(A is finite) $\leq \sum_{a \in A} \exp(s \sum_{i=1}^{m} \sigma_i a_i)$

(Easy Calculation) $\leq \sum_{a \in A} \exp(s^2 \sum_{i=1}^{m} a_i^2/2)$

$\leq \sum_{a \in A} \exp(\sum_{i=1}^{m} \sigma_i a_i)$

Therefore, $\sup_{a \in A} (\sum_{i=1}^{m} \sigma_i a_i) \leq \frac{\log |A|}{s} + \frac{sR^2}{2}$, and the right hand side is minimized when $s = \frac{\sqrt{2 \log |A|}}{R} \Rightarrow \sup_{a \in A} (\sum_{i=1}^{m} \frac{1}{m} \sigma_i a_i) \leq R \sqrt{2 \log |A|}$. 

**Corollary 1.** If $\text{VCdim}(H) = d$, then $\forall h$, $|\text{err}_s(h) - \text{err}(h)| \leq \epsilon$ if $m \geq \frac{16}{\epsilon^2} (d \log(\frac{1}{\epsilon}) + \log(\frac{1}{\delta}))$

Proof. For this proof we need to use the fact that $C[2m] = O((2m)^d) \leq \frac{(2m)^d}{d}$. This is known as Sauer’s lemma. A proof can be found in the book by Shai Shalev-Shwartz and Shai Ben-David. Applying Theorem 2, we get that $m \geq \frac{8}{\epsilon^2} \log(\frac{2}{\delta}) \geq \frac{8}{\epsilon^2} (\frac{2m}{d})^d = \frac{8d}{\epsilon^2} \log\left(\frac{2m}{d}\right) + \frac{8}{\epsilon^2} \log\left(\frac{2}{\delta}\right)$

$\Rightarrow m \geq \frac{16}{\epsilon^2} (d \log(\frac{1}{\delta}) + \log(\frac{1}{\delta}))$

The above bound can be improved to $m \geq \frac{32}{\epsilon^2} (d + \log(\frac{1}{\delta}))$. This can be done via an argument known as chaining that yields a tighter bound on the Rademacher complexity.

**Additional Readings**


\[2\text{see Chapter 27.2 in the book by Shai Shalev-Shwartz and Shai Ben-David.}\]