STAT 598Y Statistical Learning Theory **Instructor: Jian Zhang**

Lecture 6: Generalization Bound for Infinite Function Class

We introduce the generalization error bound which utilizes the growth function of $\mathcal{H}$ or VC dimension of $\mathcal{H}$ instead of the naive cardinality $|\mathcal{H}|$.

Theorem 6-1 (Vapnik-Chervonenkis). *For any $\delta > 0$, with probability at least $1 - \delta$,*

$$\forall h \in \mathcal{H},\ R(h) \leq \hat{R}_n(h) + 2\sqrt{\frac{2\log G_{\mathcal{H}}(2n) + 2\log\frac{2}{\delta}}{n}}.$$

The proof of Theorem 6-1 utilizes a technique called *symmetrization*. For notational simplicty, we will use

$$\mathbb{P}f = \mathbb{E}[f(X,Y)], \quad \mathbb{P}_n f = \frac{1}{n}\sum_{i=1}^{n} f(x_i, y_i).$$

Here $f(X,Y)$ can be thought as $\ell(Y, h(X))$. Also we define $Z_i = (X_i, Y_i)$. The key idea is to upper bound the true risk by an estimate from an independent sample, which is often known as the "ghost" sample. We use $Z_1', \ldots, Z_n'$ to denote the ghost sample and

$$\mathbb{P}_n^{'} f = \frac{1}{n}\sum_{i=1}^{n} f(x_i', y_i').$$

Then we could project the functions in $\mathcal{H}$ onto this double sample and apply the union bound with the help of the growth function $\mathcal{G}_{\mathcal{H}}(.)$ of $\mathcal{H}$.

Lemma (Symmetrization). *For any $t > 0$ such that $t \geq \sqrt{2/n}$, we have*

$$P\left(\sup_{f\in\mathcal{F}} |\mathbb{P}f - \mathbb{P}_n f| \geq t\right) \leq 2P\left(\sup_{h\in\mathcal{H}} \left|\mathbb{P}_n^{'} f - \mathbb{P}_n f\right| \geq t/2\right).$$

Proof. Let f_n be the function achieving the supremum. By triangle inequality we have

$$I\left(|\mathbb{P}f_n - \mathbb{P}_n f_n| > t\right) I\left(|\mathbb{P}f_n - \mathbb{P}_n' f_n| < t/2\right) \quad \leq \quad I\left(|\mathbb{P}_n' f_n - \mathbb{P}_n f_n| > t/2\right).$$

Taking expecation with respect to the ghost sample we have

$$I\left(|\mathbb{P}f_n - \mathbb{P}_n f_n| > t\right) P_{D_n'}\left(|\mathbb{P}f_n - \mathbb{P}_n' f_n| < t/2\right) \quad \leq \quad P_{D_n'}\left(|\mathbb{P}_n' f_n - \mathbb{P}_n f_n| > t/2\right).$$

By Chebyshev's inequality, we have

$$P_{D_n'}\left(|\mathbb{P}f_n - \mathbb{P}_n' f_n| < t/2\right) \leq \frac{4\mathbb{V}[f_n]}{nt^2} \leq \frac{1}{nt^2}$$

where $\mathbb{V}[f_n] \leq 1/4$ because f_n is a random variable (function of the sample) whose value is between 0 and 1. Hence

$$I\left(|\mathbb{P}f_n - \mathbb{P}_n f_n| > t\right)\left(1 - \frac{1}{nt^2}\right) \quad \leq \quad P_{D_n'}\left(|\mathbb{P}_n' f_n - \mathbb{P}_n f_n| > t/2\right).$$

Taking expecation with respect to the original sample and utilizing the fact that $t \geq \sqrt{2/n}$ we obtain the result. Note the same result holds if we remove the absolute operator. □

The symmetrization lemma replaces the true risk by an average over the ghost sample. As a result, the RHS only depends on the project of the function class $\mathcal{F}$ on the double sample $\mathcal{F}_{D_n, D_n'}$.

PROOF OF THEOREM 6-1:

Let $\mathcal{F} = \{f : f(x,y) = \ell(y, h(x)), h \in \mathcal{H}\}$. First note that $G_{\mathcal{H}}(n) = G_{\mathcal{F}}(n)$.

$$\begin{aligned}
P\left(\sup_{h\in\mathcal{H}} R(h) - \hat{R}_n(h)) \geq \epsilon\right) &= P\left(\sup_{f\in\mathcal{F}}(\mathbb{P}f - \mathbb{P}_n f) \geq \epsilon\right) \\
&\leq 2P\left(\sup_{f\in\mathcal{F}}(\mathbb{P}'_n f - \mathbb{P}_n f) \geq \epsilon/2\right) \\
&= 2P\left(\sup_{f\in\mathcal{F}_{D_n,D'_n}}(\mathbb{P}'_n f - \mathbb{P}_n f) \geq \epsilon/2\right) \\
&\leq 2G_{\mathcal{F}}(2n)P\left((\mathbb{P}'_n f - \mathbb{P}_n f) \geq \epsilon/2\right) \\
&\leq 2G_{\mathcal{F}}(2n)\exp(-n\epsilon^2/8).
\end{aligned}$$

The last inequality is by the Hoeffding's inequality since $P(P'_n f - P_n f \geq t) \leq \exp(-nt^2/2)$. Setting $\delta = 2G_{\mathcal{F}}(2n)\exp(-n\epsilon^2/8)$ we have the claimed result. □

Note that in the case of a finite function class, we have $|\mathcal{H}| = m$ and $G_{\mathcal{H}}(2n) \leq m$. So except for the constants, the bound is at least as good as the one before. Furthermore, if the VC dimension of $\mathcal{H}$ is $d_{VC}(\mathcal{H}) \leq n$, we can apply the Sauer's lemma to obtain the following result:

Corollary 6-2. *For any $\delta > 0$, with probability at least $1-\delta$,*

$$\forall h \in \mathcal{H},\ R(h) \leq \hat{R}_n(h) + 2\sqrt{\frac{2d_{VC}(\mathcal{H})\log\frac{2en}{d_{VC}(\mathcal{H})} + 2\log\frac{2}{\delta}}{n}}.$$

Note that in order for the result to be meaningful in Theorem 6-1, we require the $d_{VC}(\mathcal{H})$ to be finite. A class of functions whose VC dimension is finite is called a *VC class*. We can also utilize this result to obtain a bound on the expected risk $\mathbb{E}[R(\hat{h}_n)]$, where $\hat{h}_n$ is the empirical minimizer. Since

$$R(\hat{h}_n) - \inf_{h\in\mathcal{H}} R(h) \leq 2\sup_{h\in\mathcal{H}}\left|R(h) - \hat{R}_n(h)\right|,$$

we have

$$\begin{aligned}
P\left(R(\hat{h}_n) - \inf_{h\in\mathcal{H}} R(h) \geq \epsilon\right) &\leq P\left(\sup_{h\in\mathcal{H}}\left|R(h) - \hat{R}_n(h)\right| \geq \epsilon/2\right) \\
&\leq 4G_{\mathcal{H}}(2n)\exp(-n\epsilon^2/32).
\end{aligned}$$

Define a nonnegative random variable $Z = R(\hat{h}_n) - \inf_{h\in\mathcal{H}} R(h)$, and we have $P(Z \geq \epsilon) \leq 4G_{\mathcal{H}}(2n)\exp(-n\epsilon^2/32)$. Thus

$$\begin{aligned}
\mathbb{E}[Z^2] &= \int_0^\infty P(Z^2 \geq t)dt \\
&= \int_0^u P(Z^2 \geq t)dt + \int_u^\infty P(Z^2 \geq t)dt \\
&\leq u + \int_u^\infty 4G_{\mathcal{H}}(2n)\exp(-nt/32)dt \\
&= u + \frac{32G_{\mathcal{H}}(2n)}{n}\exp\left(-\frac{nu}{32}\right).
\end{aligned}$$

Minimizing the RHS with respect to u we have $u = 32G_{\mathcal{H}}(2n)/n$. Plugging in we have $\mathbb{E}[Z^2] \leq \frac{32\log(4G_{\mathcal{H}}(2n))}{n}$. By the Cauchy-Schwarz inequality we have

$$\mathbb{E}[R(\hat{h}_n)] - \inf_{h\in\mathcal{H}} R(h) = \mathbb{E}[Z] \leq \sqrt{\mathbb{E}[Z^2]} \leq O\left(\frac{\log G_{\mathcal{H}}(2n)}{n}\right).$$

So if the growth function is only polynomially increasing as a function of n, then obviously we have $\mathbb{E}[R(\hat{h}_n)] - \inf_{h \in \mathcal{H}} R(h) \to 0$, i.e. the expected risk will converge to the minimum risk within the function class $\mathcal{H}$.