Machine Learning Theory (CS 6783)

Lecture 12: Covering number, Fat-shattering, Rademacher and Supervised Learnability

1 Recap

1. Covering: \( V \) is an \( \ell_p \)-cover of \( \mathcal{F} \) on \( x_1, \ldots, x_n \) at scale \( \beta \) if
\[
\forall f \in \mathcal{F}, \exists v \in V \text{ s.t. } \left( \frac{1}{n} \sum_{t=1}^{n} |f(x_t) - v[t]|^p \right)^{1/p} \leq \beta
\]
\[\mathcal{N}_p(\mathcal{F}, \beta; x_1, \ldots, x_n) = \min\{|V| : V \text{ is an } \ell_p \text{-cover of } \mathcal{F} \text{ on } x_1, \ldots, x_n \text{ at scale } \beta\}\]

2. \[\mathbb{E}_S \left[ L_D(\hat{y}_{\text{erm}}) - \inf_{f \in \mathcal{F}} L_D(f) \right] \leq 2 \mathbb{E}_S \left[ \hat{R}_S(\mathcal{F}) \right] \leq 2 \inf_{\beta > 0} \left\{ \beta + \sqrt{\frac{\log N_1(\mathcal{F}, \beta; x_1, \ldots, x_n)}{n}} \right\}\]

3. \[\hat{R}_S(\mathcal{F}) \leq \hat{D}_S(\mathcal{F}) := \inf_{\alpha > 0} \left\{ 4\alpha + 12 \int_{\alpha}^{1} \sqrt{\log N_2(\mathcal{F}, \beta; x_1, \ldots, x_n)} d\beta \right\}\]

Also, \( \hat{R}_S(\mathcal{F}) \geq \tilde{\Omega} \left( \hat{D}_S(\mathcal{F}) \right) \)

2 Fat Shattering Dimension

**Definition 1.** We say that \( \mathcal{F} \) shatters \( x_1, \ldots, x_n \) at scale \( \gamma \), if there exists witness \( s_1, \ldots, s_n \) such that, for every \( \epsilon \in \{\pm 1\}^n \), there exists \( f_\epsilon \in \mathcal{F} \) such that
\[
\forall t \in [n], \quad \epsilon_t \cdot (f_\epsilon(x_t) - s_t) \geq \gamma / 2
\]

Further \[\text{fat}_\gamma(\mathcal{F}) = \max\{n : \exists x_1, \ldots, x_n \in \mathcal{X} \text{ s.t. } \mathcal{F} \text{-shatters } x_1, \ldots, x_n\}\]
**Theorem 1.** For any $\mathcal{F} \subseteq [-1, 1]^X$ and any $\gamma \in (0, 1)$

$$
\mathcal{N}_2(\mathcal{F}, \gamma, n) \leq \left( \frac{2}{\gamma} \right)^{K \text{ fat}_{c\gamma}(\mathcal{F})}
$$

where in the above $c$ and $K$ are universal constants.

Using the above with the dudley chaining bounds we get,

$$
\mathcal{D}_S(\mathcal{F}) \leq \inf_{\alpha \geq 0} \left\{ 4\alpha + \frac{12}{\sqrt{n}} \int_0^1 \sqrt{K \text{ fat}_{c\delta}(\mathcal{F})} \log \left( \frac{2}{\delta} \right) d\delta \right\}
$$

Thus bound on fat-shattering dimension leads to bound on Rademacher complexity.

**Binary function class** For any $\delta \in [0, 1)$, and any $c \leq 1$, $\text{fat}_{c\delta}(\mathcal{F}) = \text{fat}_0(\mathcal{F}) = \text{VC}(\mathcal{F})$ we can conclude that

$$
\mathcal{V}^\text{stat}_n(\mathcal{F}) \leq \mathcal{R}_n(\mathcal{F}) \leq \sqrt{\text{VC}(\mathcal{F})} n.
$$

**Linear Predictors** Let $\mathcal{X} = \{ x : \| x \|_2 \leq 1 \}$ and let $\mathcal{F} = \{ x \mapsto f^\top x : \| f \|_2 \leq 1 \}$.

1. $\text{fat}_\gamma(\mathcal{F}) \geq [4\gamma^{-2}]$

   For all $i \in [d]$, let $x_i = e_i$ and let $s_i = 0$. Given $\epsilon \in \{-1\}^d$, consider the vector $f$ such that $f[i] = \epsilon_i \gamma/2$. Clearly $f$, $\gamma$-shatters these set of $d$ points. Now for $\| f \|_2 \leq 1$, we need that $\sum_{i=1}^d f^2[i] = d\gamma^2/4 \leq 1$. This implies that $d \leq 4\gamma^{-2}$. Thus we can provide $4/\gamma^2$ points that can be $\gamma$-shattered.

2. $\text{fat}_\gamma(\mathcal{F}) \leq 4\gamma^{-2}$

   Typically uses Maurey’s theorem but we will take a different route in just a bit.

### 2.1 Back to Rademacher

Let us define the worst case Rademacher complexity as follows :

$$
\mathcal{R}_n(\mathcal{F}) = \sup_{x_1, \ldots, x_n} \frac{1}{n} \mathbb{E}_\epsilon \left[ \sup_{f \in \mathcal{F}} \sum_{t=1}^n \epsilon_t f(x_t) \right]
$$

We have the following lower bound on the worst case Rademacher complexity.

**Claim 2.**

$$
\mathcal{R}_n(\mathcal{F}) \geq \sup \{ \gamma/2 : \text{fat}_\gamma(\mathcal{F}) > n \}
$$

**Proof.** Think about Rademacher complexity on shattered points. \qed

The claim above is the same as saying (converse) $\text{fat}_\gamma \leq \min \{ n : \mathcal{R}_n(\mathcal{F}) \leq \gamma/2 \}$. Using this for linear class example, since we know that $\mathcal{R}_n(\mathcal{F}) \leq \frac{1}{\sqrt{n}}$, we can conclude that for the linear class, $\text{fat}_\gamma \leq \min \{ n : \mathcal{R}_n(\mathcal{F}) \leq \gamma/2 \} \leq \min \{ n : \frac{1}{\sqrt{n}} \leq \gamma/2 \} \leq \lceil \frac{4}{\gamma^2} \rceil$.
Using a more refined argument, the claim above can be improved, it can be shown that for any $\gamma > R_n(F)$,
\[
\text{fat}_\gamma(F) \leq \frac{8nR_n^2(F)}{\gamma^2}
\]
from this we can conclude that
\[
\hat{R}_S(F) \geq \Omega\left(\inf_{\delta \geq 0} \left\{ 4\alpha + \frac{12}{\sqrt{n}} \int_0^1 \sqrt{K \text{fat}_\alpha(F) \log(2/\delta) \delta} \right\} \right)
\]

3 Lower Bounds on Supervised Learning for $Y \subset \mathbb{R}$

Basic idea: To show lower bound, we pick $k \cdot n$ points $x_1, \ldots, x_{kn}$ and signs $\epsilon_1, \ldots, \epsilon_{kn}$. The signs are not revealed to the learner. We use the uniform distribution over the $kn$ pairs of instances as the distribution $D$. That is $D = \text{Unif}\{(x_1, \epsilon_1), \ldots, (x_{kn}, \epsilon_{kn})\}$. Learner can even know this fact, only learner does not get to see the $\epsilon_i$’s before hand. Now we sample $n$ points from this distribution and provide this to the learner. Clearly the learner sees at most $n$ labels and so on the the remaining $kn - n$ points learner has no way to predict anything meaningful. The rest is simply massaging the math.

We shall consider the absolute loss $\ell(y, y') = |y - y'|$. However similar analysis can be extended to other commonly used supervised learning losses (called margin losses) like all $\ell_p$ losses, logistic loss, hinge loss etc.

Lemma 3. For any class $F \subset [-1, 1]^X$ and for any $k \in \mathbb{N}$,
\[
\nu_n^{\text{proper}}(F) \geq R_{kn} - \frac{1}{k} R_n(F) \quad \text{and} \quad \nu_n^{\text{improper}}(F) \geq R_{kn} - \frac{1}{k}
\]

Proof.
\[
\inf_{\hat{y}} \sup_{D} \mathbb{E}_S \left[ L_D(\hat{y}) - \inf_{f \in F} L_D(f) \right] \\
\geq \inf_{\hat{y}} \sup_{x_1, \ldots, x_{kn}} \mathbb{E}_{\hat{y} \sim \text{Unif}\{(x_1, \epsilon_1), \ldots, (x_{kn}, \epsilon_{kn})\}} \left[ \frac{1}{kn} \sum_{t=1}^{kn} |\hat{y}_S(x_t) - \epsilon_t| - \inf_{f \in F} \frac{1}{kn} \sum_{t=1}^{kn} |f(x_t) - \epsilon_t| \right] \\
\geq \sup_{x_1, \ldots, x_{kn}} \inf_{\hat{y}} \mathbb{E}_{\hat{y} \sim \text{Unif}\{(x_1, \epsilon_1), \ldots, (x_{kn}, \epsilon_{kn})\}} \left[ \frac{1}{kn} \sum_{t=1}^{kn} |\hat{y}_S(x_t) - \epsilon_t| - \inf_{f \in F} \frac{1}{kn} \sum_{t=1}^{kn} |f(x_t) - \epsilon_t| \right] \\
\quad \text{For any } y' \in [-1, 1], |y' - \epsilon_t| = 1 - y'\epsilon_t \text{ and so,}
\]
\[
= \sup_{x_1, \ldots, x_{kn}} \inf_{\hat{y}} \mathbb{E}_{\hat{y} \sim \text{Unif}\{(x_1, \epsilon_1), \ldots, (x_{kn}, \epsilon_{kn})\}} \left[ \frac{1}{kn} \sum_{t=1}^{kn} -\epsilon_t \hat{y}_S(x_t) - \inf_{f \in F} \frac{1}{kn} \sum_{t=1}^{kn} -\epsilon_t f(x_t) \right] \\
= \sup_{x_1, \ldots, x_{kn}} \left\{ \inf_{\hat{y}} \mathbb{E}_{e} \left[ \frac{1}{kn} \sum_{t=1}^{kn} -\epsilon_t \hat{y}_S(x_t) \right] - \mathbb{E}_e \left[ \inf_{f \in F} \frac{1}{kn} \sum_{t=1}^{kn} -\epsilon_t f(x_t) \right] \right\} \\
= \sup_{x_1, \ldots, x_{kn}} \left\{ \mathbb{E}_e \left[ \sup_{f \in F} \frac{1}{kn} \sum_{t=1}^{kn} \epsilon_t f(x_t) \right] - \sup_{\hat{y}} \mathbb{E}_S \mathbb{E}_{\hat{y}} \left[ \frac{1}{kn} \sum_{t=1}^{kn} \epsilon_t \hat{y}_S(x_t) \right] \right\}
\]
Now define \( J \subset [2n] \) as, \( J_S = \{ i : (x_i, \epsilon_i) \in S \} \). Notice that for any \( i \in J_S \), because \( \hat{y}_S \) is only a function of sample \( S \), we have \( \mathbb{E}_S [ \mathbb{E}_{\epsilon_i} [ \epsilon_i \hat{y}_S(x_i) ] ] = \mathbb{E}_S [ \mathbb{E}_{\epsilon_i} [ \epsilon_i ] \hat{y}_S(x_i) ] = 0 \). Hence:

\[
\mathcal{V}_n^{\text{stat}}(\mathcal{F}) \geq \sup_{x_1, \ldots, x_{kn}} \left\{ \mathbb{E}_e \left[ \sup_{f \in \mathcal{F}} \frac{1}{kn} \sum_{t=1}^{kn} \epsilon_t f(x_t) \right] - \frac{1}{kn} \sup_{\hat{y}} \mathbb{E}_S \mathbb{E}_e \left[ \sum_{t \in J} \epsilon_t \hat{y}(x_t) \right] \right\}
\]

\[
\geq \sup_{x_1, \ldots, x_{kn}} \mathbb{E}_e \left[ \frac{1}{kn} \sum_{t=1}^{kn} \epsilon_t f(x_t) \right] - \frac{1}{kn} \sup_{\hat{y}} \mathbb{E}_S \mathbb{E}_e \left[ \sum_{t \in J} \epsilon_t \hat{y}(x_t) \right]
\]

\[
= \mathcal{R}_{kn}(\mathcal{F}) - \frac{1}{kn} \sup_{\hat{y}} \mathbb{E}_e \left[ \sum_{t=1}^{n} \epsilon_t \hat{y}(x_t) \right]
\]

Now if we consider minimax rates with respect to only proper learning algorithms, that is \( \hat{y}_S \in \mathcal{F} \), then

\[
\mathcal{V}_n^{\text{stat}}(\mathcal{F}) \geq \mathcal{R}_{kn}(\mathcal{F}) - \frac{1}{kn} \sup_{\hat{y}} \mathbb{E}_e \left[ \sum_{t=1}^{n} \epsilon_t \hat{y}(x_t) \right]
\]

\[
\geq \mathcal{R}_{kn}(\mathcal{F}) - \frac{1}{k} \mathcal{R}_n(\mathcal{F})
\]

On the other hand if we consider improper learning algorithms as well, then

\[
\mathcal{V}_n^{\text{stat}}(\mathcal{F}) \geq \mathcal{R}_{kn}(\mathcal{F}) - \frac{1}{kn} \sup_{\hat{y}} \mathbb{E}_e \left[ \sum_{t=1}^{n} \epsilon_t \hat{y}(x_t) \right] \geq \mathcal{R}_{kn}(\mathcal{F}) - \frac{1}{k}
\]

Using \( k = 2 \), in the above, we get that for proper learning algorithms, \( \mathcal{V}_n^{\text{stat}}(\mathcal{F}) \geq \mathcal{R}_{2n}(\mathcal{F}) - \frac{1}{2} \mathcal{R}_n(\mathcal{F}) \). If \( \mathcal{R}_n(\mathcal{F}) = \Theta(n^{-p}) \) for some \( p \geq 2 \) then, from this we conclude that if we consider minimax rate for proper learning,

\[
\mathcal{V}_n^{\text{stat}}(\mathcal{F}) \geq 0.29 \mathcal{R}_{2n}(\mathcal{F})
\]

On the other hand if we consider improper learning as well, if \( \mathcal{R}_n(\mathcal{F}) = \Omega(n^{-1/p}) \) then picking \( k = 2n^{1/(p-1)} \), in the lower bound above for improper learning we can conclude that,

\[
\mathcal{V}_n^{\text{stat}}(\mathcal{F}) \geq \Omega \left( n^{-\frac{1}{p-1}} \right)
\]

4 Putting It All Together

**Theorem 4.** For any real valued hypothesis class \( \mathcal{F} \), and supervised statistical learning problem with absolute loss (also for squared loss, logistic loss, . . . ), the following are equivalent:

1. \( \mathcal{F} \) is uniformly learnable (\( \mathcal{V}_n^{\text{stat}}(\mathcal{F}) \to 0 \))
2. \( \mathcal{R}_n(\mathcal{F}) \to 0 \)
3. \( \mathcal{D}_n(\mathcal{F}) \to 0 \)
4. \( \forall \gamma > 0, \text{fat}_{\gamma} < \infty \)
Summary:

1. We have a crisp certificate for learnability for real valued supervised learning problems. Rates are tight for absolute loss, hinge loss and zero-one loss.

2. Any one of Rademacher complexity, covering numbers or fat-shattering dimension can provide to within log factors the optimal rates.