Lecture 24: Lower Bound Sketch + Margin Theory Sketch

Lecturer: Jacob Abernethy *Scribes: Felipe Lagos, Majid Ahadi*

24.1 Lower bound on the generalization error

So far, we have shown that if a class $\mathcal{H}$ with VCdim $= d$, then ERM guarantees with probability at least $1-\delta$, the following upper bound,

$$R(\hat{h}_S) - R(h^*) \leq c\sqrt{\frac{d\log(h)}{m}} + \sqrt{\frac{\log(2/\delta)}{2m}}, \tag{24.1}$$

for all distribution $D \sim (x, y)$, $S \sim D^m$.

A lower bound can be determined by finding a 'bad' distribution for any learning algorithm. Since, the algorithm is arbitrary, it is difficult to specify that particular distribution. However, using the *probabilistic method* proof technique, it is possible to prove that there exists a distribution such that the generalization error is at least some factor $\mathcal{O}(\sqrt{d/m})$ with a constant probability. In particular, we have the following Theorem.

Theorem 24.1 *Let $\mathcal{H}$ be a hypothesis set with* VCdim $= d > 1$. *Then, for any learning algorithm, there exists a distribution D such that:*

$$\mathbb{P}_{S\sim D^m}\left(R(\hat{h}_S) - R(h^*) > \sqrt{\frac{d}{320m}}\right) \geq 1/64. \tag{24.2}$$

Here, $\hat{h}_S$ is any estimator that outputs $\hat{h}$ for S. The Theorem also states that for any learning algorithm, the sample complexity verifies,

$$m \geq \frac{d}{320\epsilon^2}. \tag{24.3}$$

The following Lemma is needed for the lower bound proof.

Lemma 24.2 *Let α be uniformly distributed in $\{\alpha_-, \alpha_+\}$, where $\alpha_+ = 1/2 + \epsilon/2$ and $\alpha_- = 1/2 - \epsilon/2$. Let D_α a distribution such that $\mathbb{P}(y = 1) = \alpha$ and $\mathbb{P}(y = 0) = 1 - \alpha$. Let $S \sim D_\alpha^m$ and let h be any estimator, $h(S) = \alpha_+$ or $h(S) = \alpha_-$ (h guesses whether it is the case α_+ or α_-). Then,*

$$\mathbb{E}_\alpha\left(\mathbb{P}_{S\sim D^m}(h(S) \neq \alpha)\right) \geq \frac{1}{4}\left(1 - \sqrt{\exp\left(-\frac{m\epsilon^2}{1-\epsilon^2}\right)}\right) := \phi(m, \epsilon). \tag{24.4}$$

Takeaway: If $m < \mathcal{O}\left(\frac{1}{\epsilon^2}\right)$, then there is a constant probability that $h(S)$ is wrong (in expected value). This lemma is like the 'opposite' of Hoeffding's inequality (if $m \geq 1/\epsilon^2 \log(1/\delta$, then $\mathbb{P}(\text{threshold estimator}(S) = \alpha) \geq 1 - \delta)$.

The Sketch proof of Theorem 24.1 is the following.

Proof: (**Sketch**) Let $S \subseteq X$ be a shattered set, with $x_i \sim \text{Unif}(S)$. Let $\sigma_1, \ldots, \sigma_d$ be random variables, uniform in $\{-1, 1\}$. Let D_σ a distribution such that $\mathbb{P}(y = 1) = 1/2 + \epsilon/2\sigma_i$ and $\mathbb{P}(y = 0) = 1/2 - \epsilon/2\sigma_i$.

If $\mathbb{E}_\sigma(R(\hat{h}_S) - R(h^*)) > \epsilon$, then, via probabilistic method, there exists a distribution D_σ such that $R(\hat{h}_S) - R(h^*) > \epsilon$. Then, using that $S_i \approx \frac{m}{d}$ (we see each sample uniformly), we have,

$$\mathbb{E}_\sigma\left(R(\hat{h}_S) - R(h^*)\right) = \mathbb{E}_\sigma\left(\frac{1}{d}\sum_{i=1}^{d}\mathbb{I}(h_{S_i}(x_i) \neq y_i)\right) = \epsilon\frac{1}{d}\mathbb{E}_\sigma\left(\sum_{i=1}^{d}\mathbb{P}_{S_i\sim D_{\sigma_i}^{n/d}}(h_{S_i} = \sigma_i)\right) = \epsilon\phi(m/d, \epsilon)$$

If we pick $\epsilon = \sqrt{\frac{d}{m}}$, the last term is constant.

■

24.2 Margin theory sketch

In this section an upper bound on the Empirical Rademacher complexity of a class, which is comprised of linear classifiers is found.
A linear classifier is defined as:

$$h_w(\vec{x}) = sign(\vec{w} \cdot \vec{x})$$

where $\vec{w} \in \mathbb{R}^N$ and $\vec{x} \subseteq \mathbb{R}^N$. Note that the bias is embedded in $\vec{x}$. A class of linear classifiers is shown as:

$$H_{lin} = \{h_w : \vec{w} \in \mathbb{R}^N\}$$

We know that the VC-dim (H_{lin}) is equal to $N+1$, and $N > m$; therefore,

$$R(\hat{h}) - R(h^*) \leq \sqrt{\frac{N\log m}{m}}$$

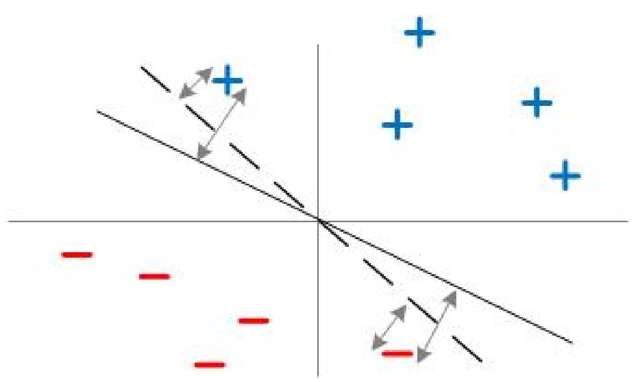
where $\sqrt{\frac{N\log m}{m}} \geq 1$; therefore this bound is not useful.
To find a tighter lower bound, consider the following class of linear classifiers with bounded norm:

$$H_\Lambda := \{h_w(\vec{x}) = \vec{w} \cdot \vec{x} : ||\vec{w}||_2 \leq \Lambda\}$$

Given a sample $S = \{(x_i, y_i) : i = 1, ..., m\}$ and $\vec{w} \in \mathbb{R}^N$, margin of the linear classifier is defined as:

$$\rho(S) = \min_{i=1,\ldots,m} \frac{y_i(\vec{w} \cdot x_i)}{||w||}$$

The key idea is that a classifier with a larger margin works better, and has a better generalization. For example, in the following figure the solid line has a larger margin; thus, it is better.



Next, Let $S = \{(x_i, y_i) : i = 1, ..., m\}$, and $||x_i|| \leq r$, where r is some constant.
Claim: Empirical Rademacher complexity of H_Λ is bounded as:

$$\hat{\mathfrak{R}}_S \leq \sqrt{\frac{r^2 \Lambda^2}{m}}$$

Proof:

$$\hat{\mathfrak{R}}_S = \mathop{\mathbb{E}}_{\sigma_1,...,\sigma_m}\left[\frac{1}{m}\sup_{h_w \in H_\Lambda}\sum_{i=1}^{m}\sigma_i h_w(x_i)\right] = \frac{1}{m}\mathop{\mathbb{E}}_{\sigma_1,...,\sigma_m}\left[\sup_{w:||w||\leq\Lambda}\left(\sum_{i=1}^{m}\sigma_i x_i\right)\cdot w\right]$$

where $\sigma_1, ..., \sigma_m$ are Rademacher random variables. Using the Cauchy-Schwarz inequality:

$$\leq \frac{1}{m}\mathop{\mathbb{E}}_{\sigma_1,...,\sigma_m}\left[\sup_{w:||w||\leq\Lambda}||w||_2\left|\left|\sum_{i=1}^{m}\sigma_i x_i\right|\right|_2\right]$$

where $\sup\limits_{w:||w||\leq\Lambda}||w||_2 \leq \Lambda$. Using the Jensen's inequality:

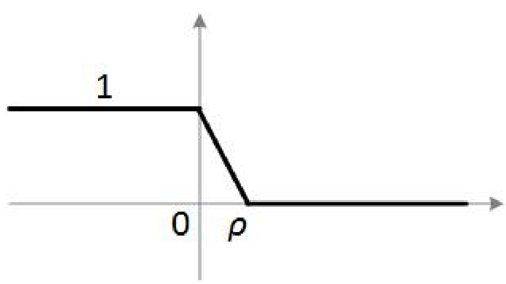
$$\leq \sqrt{\frac{\Lambda}{m}\mathop{\mathbb{E}}_{\sigma_1,...,\sigma_m}\left[\left|\left|\sum_{i=1}^{m}\sigma_i x_i\right|\right|^2\right]} = \frac{\Lambda}{m}\sqrt{\mathop{\mathbb{E}}_{\sigma_1,...,\sigma_m}\left[\sum_{i,j}\sigma_i\sigma_j x_i x_j\right]} = \frac{\Lambda}{m}\sqrt{\sum_{i=1}^{m}||x_i||^2} \leq \frac{\Lambda}{m}\sqrt{mr^2} = \sqrt{\frac{r^2\Lambda^2}{m}}$$

■

Note that the bound can be tuned by increasing r or equivalently decreasing Λ.
To somehow include the sign function we use a loss function, and find an upper bound on $\hat{\mathfrak{R}}_S(\ell \cdot H_\Lambda)$. Consider the following function:

$$\phi(z) = \begin{cases} 1 & z \leq 0 \\ 0 & z \geq \rho \\ 1 - \frac{z}{\rho} & 0 \leq z \leq \rho \end{cases}$$

which looks like:



In fact, this function makes us to pay for low confidence. If we use $\ell(y, \hat{y}) = \phi_\rho(y\hat{y})$, where $y \in \{-1, 1\}$, it can be proven that:

$$\hat{\mathfrak{R}}_S(\ell \cdot H_\Lambda) \leq \frac{1}{\rho}\hat{\mathfrak{R}}_S(H_\Lambda) \leq \frac{1}{\rho}\sqrt{\frac{r^2\Lambda^2}{m}}$$

For more information and the proof look at the Lemma 4.2 Talagrand's Lemma in the book.