CSC 588: Machine learning theory **Spring 2021**

Lecture 13: More on weak learnability; Margin-based generalization bounds

Lecturer: Chicheng Zhang *Scribe: Ryan Sullivant*

1 Weak learnability implies "linear separability" - a nonconstructive view

Recall: $\mathcal{A}$ is a γ-weak PAC learner for $\mathcal{H}$ if there is a sample complexity function $f_S : (0,1) \to \mathbb{N}$ such that for any $\delta > 0$ and any distribution $\mathcal{D}$ realizable by $\mathcal{H}$ with m i.i.d. training examples, $\mathcal{A}$ produces a classifier f such that with probability $1 - \delta$

$$\text{err}(f, \mathcal{D}) \leq \frac{1}{2} - \gamma$$

Now suppose we have a γ-weak learner $\mathcal{A}$ which returns classifiers from a base hypothesis class $\mathcal{B}$. Then, for any $h^* \in \mathcal{H}$, and any distribution $\mathcal{D}_X$ over $\mathcal{X}$, there is an $f \in \mathcal{B}$ (returned by $\mathcal{A}$ with nonzero probability) such that

$$\mathbb{E}_{x \sim \mathcal{D}_X} I(f(x) \neq h^*(x)) \leq \frac{1}{2} - \gamma.$$

Informally, the above statement says that, fix any $h^\star$ in $\mathcal{H}$; then for any distribution over the unlabeled examples $\mathcal{D}_X$, we can find a classifier from the base class that nontrivially correlates with $h^\star$ with respect to $\mathcal{D}_X$ (recall that a classifier that performs random guessing would have an error of exactly $\frac{1}{2}$ with respect to $h^\star$). AdaBoost gives a way to generate a set of base classifiers and weights whose induced weighted majority vote can express $h^\star$ exactly (after taking signs), summarized below:

Claim 1. *Suppose we have a γ-weak learner $\mathcal{A}$ for $\mathcal{H}$ which returns classifiers from a base hypothesis class $\mathcal{B}$. Then, for any $h^\star \in \mathcal{H}$, there is a distribution $\mathcal{D}_\mathcal{B}$ over $\mathcal{B}$ such that*

$$h^*(x) = \text{sign}(\sum_{f \in \mathcal{B}} \mathcal{D}_\mathcal{B}(f) f(x))$$

Remark 2. *This is a statement about the expressiveness of the base class $\mathcal{B}$. Note that AdaBoost gives an algorithmic proof of this.*

We now provide another perspective of the claim by giving a more direct (nonconstructive) proof of it.

Proof. Recall from the above discussion that for any $h^* \in \mathcal{H}$, and any distribution $\mathcal{D}_X$ over $\mathcal{X}$, there is an $f \in \mathcal{B}$ (returned by $\mathcal{A}$ with nonzero probability) such that

$$\mathbb{E}_{x \sim \mathcal{D}_X} I(f(x) \neq h^*(x)) \leq \frac{1}{2} - \gamma.$$

The statement can be rewritten using max and min as follows:

$$\max_{\mathcal{D}_X} \min_{f \in \mathcal{B}} \mathbb{E}_{x \sim \mathcal{D}_X} I(f(x) \neq h^*(x)) \leq \frac{1}{2} - \gamma$$

This is equivalent to

$$\max_{\mathcal{D}_X} \min_{f \in \mathcal{B}} \mathbb{E}_{x \sim \mathcal{D}_X} \left[\frac{1 - f(x)h^*(x)}{2} \right] \leq \frac{1}{2} - \gamma$$

Which is equivalent to

$$\min_{\mathcal{D}_X} \max_{f \in \mathcal{B}} \mathbb{E}_{x \sim \mathcal{D}_X} f(x)h^*(x) \geq 2\gamma$$

Letting $\Delta(B)$ be the set of all distributions support on B, we can rewrite the left hand side of this inequality as

$$\min_{\mathcal{D}_X} \max_{\mathcal{D}_\mathcal{B} \in \Delta(B)} \mathbb{E}_{f \sim \mathcal{D}_\mathcal{B}} \mathbb{E}_{x \sim \mathcal{D}_X} f(x) h^*(x) \tag{1}$$

We will use the Von Neumann Minimax Theorem to obtain the required bound.

Theorem 3 (Von Neumann)**.** *Let* $\Delta^d = \{(\nu_1, \ldots, \nu_d) \mid \forall i, \nu_i \geq 0 \text{ and } \sum_i \nu_i = 1\}$. *Given* $A \in \mathbb{R}^{m \times n}$, *the following holds*

$$\min_{p \in \Delta^m} \max_{q \in \Delta^n} p^\top A q = \max_{q \in \Delta^n} \min_{p \in \Delta^m} p^\top A q$$

Please see below the end of the proof for some illuminating examples of what this theorem says. For now, we will complete the proof. We will apply Von Neumann's theorem to Equation 1 by letting $p = \mathcal{D}_X$ and $q = \mathcal{D}_\mathcal{B}$, so that Equation 1 becomes

$$\max_{\mathcal{D}_\mathcal{B}} \min_{\mathcal{D}_X} \mathbb{E}_{x \sim \mathcal{D}_X} \mathbb{E}_{f \sim \mathcal{D}_\mathcal{B}} f(x) h^*(x)$$

This implies that

$$\max_{\mathcal{D}_\mathcal{B}} \min_{\mathcal{D}_X} \mathbb{E}_{x \sim \mathcal{D}_X} \mathbb{E}_{f \sim \mathcal{D}_\mathcal{B}} f(x) h^*(x) \geq 2\gamma$$

Equivalently,

$$\max_{\mathcal{D}_\mathcal{B}} \min_{x} \mathbb{E}_{f \sim \mathcal{D}_\mathcal{B}} f(x) h^*(x) \geq 2\gamma$$

Changing back to quantifiers this equation says:

$$\text{there is a } \mathcal{D}_\mathcal{B} \text{ such that for all } x \in X \sum_{f \in \mathcal{D}_\mathcal{B}} \mathcal{D}_\mathcal{B}(f) f(x) h^*(x) \geq 2\gamma > 0$$

□

Remark 4. *Suppose that* $\mathcal{B}$ *is finite and for each* $x \in X$ *define* $\tilde{x} = (f(x))_{f \in \mathcal{B}} \in \{\pm 1\}^{|\mathcal{B}|}$. *Then,* $h^*(x) = \text{sign}(\sum_{f \in \mathcal{B}} \mathcal{D}_\mathcal{B}(f) f(x))$ *implies that,* $(\tilde{x}, h^*(x))_{x \in X}$ *is linearly separable by*

$$w = (\mathcal{D}_\mathcal{B}(f))_{f \in \mathcal{B}} \in \mathbb{R}^{|\mathcal{B}|}$$

1.1 Examples for Von Neumann's theorem

Suppose that Alice plays against Bob in a (zero-sum) game of Rock-Paper-Scissors. Suppose that the payoff matrix A (the loss of Alice = the gain of Bob) of the game is

$A =$

Alice \ Bob	R	P	S
R	0	1	-1
P	-1	0	1
R	1	-1	0

We will consider a couple different protocols.

Protocol 1

- Alice goes first and chooses row i
- Bob reponds by choosing col j after seeing i
- Alice suffers loss of $A_{i,j}$

If Alice and Bob behave optimally, what is Alice's payoff? Given, information about i, Bob can always find j such that $A_{i,j} = 1$, i.e. given i, $\max_j A_{i,j} = 1$. As this holds for any i,

$$\min_i \max_j A_{i,j} = 1$$

Hence, Alice always suffers a loss of 1 in this protocol.

Protocol 2

- Bob goes first and chooses col j
- Alice reponds by choosing row i after seeing j
- Alice suffers loss of $A_{i,j}$

If Alice and Bob behave optimally, what is Alice's payoff? Given, information about j, Alice can always find i such that $A_{i,j} = -1$, i.e. given j, $\min_i A_{i,j} = -1$. As this holds for any j,

$$\max_j \min_i A_{i,j} = -1$$

And it follows that Alice receives a loss of -1 in this protocol.

Chicheng notes: the above two protocols are also sometimes called "Stackelberg games" where two players move sequentially. In summary, with the ability to take actions at later stages, Alice gains an advantage and can suffer a lower loss. This is true in general, in that for any matrix A, $\max_j \min_i A_{i,j} \leq \min_i \max_j A_{i,j}$. Can you see why?

Now consider the following modifications to Protocol 1 and 2.

Protocol 1^*

- Alice goes first and chooses distribution $p \in \Delta^m$ over rows
- Bob reponds by choosing distribution $q \in \Delta^n$ over cols after seeing p
- Alice suffers loss of $p^\top Aq = \mathbb{E}_{i\sim p}\mathbb{E}_{j\sim q}A_{i,j}$

Similarly,

Protocol 2^*

- Bob goes first by choosing distribution $q \in \Delta^n$ over cols
- Alice responds and chooses distribution $p \in \Delta^m$ over rows after seeing q
- Alice suffers loss of $p^\top Aq = \mathbb{E}_{i\sim p}\mathbb{E}_{j\sim q}A_{i,j}$

Von Neumann's Theorem says that Alice's payoff in Protocol 1^* equals her payoff in Protocol 2^*; in other words, being allowed to delay her decision in the modified games does not give Alice an advantage.

2 Generalization error bounds for boosting

Consider choosing T in AdaBoost

- As T increases, $\text{err}(H_T, S)$ decreases as $\text{err}(H_T, S) \leq \exp(-2T\gamma^2)$
- What about $\text{err}(H_T, \mathcal{D})$? We can bound is using VC dimension. Considering

$$H_T \in \mathcal{H}_T := \left\{ \text{sign}\left(\sum_{t=1}^{T} \alpha_t h_t(x) \right) \mid \forall t, \alpha_t \in \mathbb{R} \text{ and } h_t \in \mathcal{B} \right\}$$

Then,

$$\text{err}(H_T, D) \leq \text{err}(H_T, S) + \sqrt{\frac{VC(\mathcal{H}_T)}{m}}$$

However, $VC(\mathcal{H}_T)$ depends linearly on T.

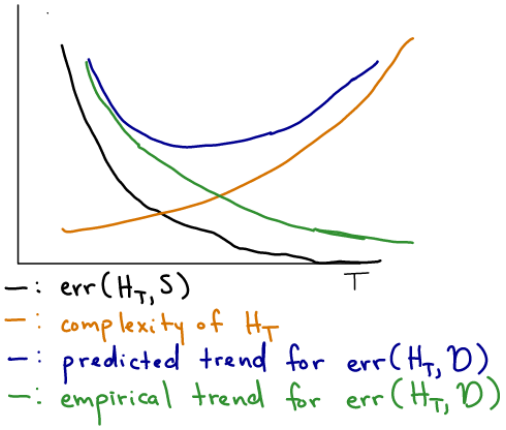

Figure 1: Graph showing showing the relationship between $\text{err}(H_T, S)$, complexity of H_T, and $\text{err}(H_T, \mathcal{D})$. The bound on $\text{err}(H_T, \mathcal{D})$ using VC dimension would predict that $\text{err}(H_T, \mathcal{D})$ eventually increases. However, in practice it can continue to decrease even after $\text{err}(H_T, S)$ reaches zero.

Because of this linear dependence, we would expect that $\text{err}(H_T, \mathcal{D})$ to increase as T increase. However, in practice this does not seem to be the case (cf. Figure 1). Empirically, there is a trend for $\text{err}(H_T, \mathcal{D})$ to decrease even once $\text{err}(H_T, S)$ reaches zero. Looking for an explanation to this phenomenom led to the *large margin theory for boosting*.

Definition 5. *For a function f and data point (x, y), the* margin of f on (x, y) *is $yf(x)$.*

Theorem 6. *Suppose that base hypothesis class $\mathcal{B}$ is finite. Let*

$$C(\mathcal{B}) = \left\{ \sum_{h \in \mathcal{B}} \alpha_h h(x) \mid \sum_{h \in \mathcal{H}} |\alpha_h| \leq 1 \right\}$$

denote the set of voting classifiers over $\mathcal{B}$. Fix margin $\theta \in [0, 1]$. Then, given i.i.d. $S \sim \mathcal{D}$ of size m, with probability $1 - \delta$, for all $f \in C(\mathcal{B})$

$$\mathbb{P}_{\mathcal{D}}(yf(x) \leq 0) \leq \mathbb{P}_S(yf(x) \leq \theta) + O\left(\frac{1}{\theta}\sqrt{\frac{\ln|\mathcal{B}|/\delta}{m}}\right)$$

Remark 7. *Note that $\mathbb{P}_{\mathcal{D}}(yf(x) \leq 0)$ is the error of* $\text{sign}(f)$*. Furthermore, notice that the error bound depends on $\ln|\mathcal{B}|$ whereas the simple bound above on $\text{err}(H_T, \mathcal{D})$ for AdaBoost depended on $\sqrt{T}$.*

We can apply this to AdaBoost since

$$\bar{f}_t = \frac{\sum_{t=1}^{T} \alpha_t h_t(x)}{\sum_{t=1}^{T} \alpha_t} \in C(\mathcal{B})$$

Then, letting $\theta = \frac{\gamma}{2}$ we have

1. $\mathbb{P}_S(yf_T \leq \frac{\gamma}{2}) \leq \exp(-T\gamma^2)$ (This requires some work and is left as an exercise for the interested.)

2. The complexity term is $O\left(\frac{1}{\theta}\sqrt{\frac{\ln|\mathcal{B}|/\delta}{m}}\right)$ does not depend on the number of iterations T.

Instead of proving the theorem in this formulation, we will prove the following more general version.

Theorem 8. *Suppose $\mathcal{D}$ is supported on $\{x \in \mathbb{R}^d \mid \|x\|_\infty \le R_\infty\} \times \{\pm 1\}$. Fix value $\theta \in [0,1]$. Suppose that S has size m and is drawn i.i.d. from $\mathcal{D}$. Then, with probability $1-\delta$, for any predictor w such that $\|w\|_1 \le B_1$*

$$\mathbb{P}_{\mathcal{D}}(y\langle w, x\rangle \le 0) \le \mathbb{P}_S(y\langle w, x\rangle \le \theta) + O\left(\frac{B_1 R_\infty}{\theta}\sqrt{\frac{\ln d/\delta}{m}}\right)$$

To see that we can use this theorem to get the previous one, fix $\mathcal{B} = \{h_1, \dots, h_N\}$. Then, let

- $\tilde{x} = (h_1(x), \dots, h_N(x))$ so that $(\tilde{x}, y) \sim \tilde{\mathcal{D}}$. Then, $\|\tilde{x}\|_\infty = 1 =: R_\infty$
- We let α play the role of w since $\sum_{i=1}^N \alpha_i h_i(x) = \langle \alpha, \tilde{x}\rangle$ and $\|\alpha_i\|_1 \le 1$ for all i, so we can set $B_1 = 1$.
- Finally, use N for d.

Remark 9. *Note that this is not a standard generalization to training error bound; the left and right hand side errors are not completely symmetric.*

We start the proof now and finish it next time

Proof. Define the *ramp loss* $\ell_{\theta,w}(x,y)$ as

$$\ell_{\theta,w}(x,y) = \varphi_\theta(y\langle w, x\rangle)$$

where

$$\varphi_\theta(z) = \begin{cases} 1 & z \le 0 \\ 1 - \frac{z}{\theta} & z \in (0,\theta) \\ 0 & z \ge \theta \end{cases}$$

See Figure 2, to see a comparison between the three losses we are considering.

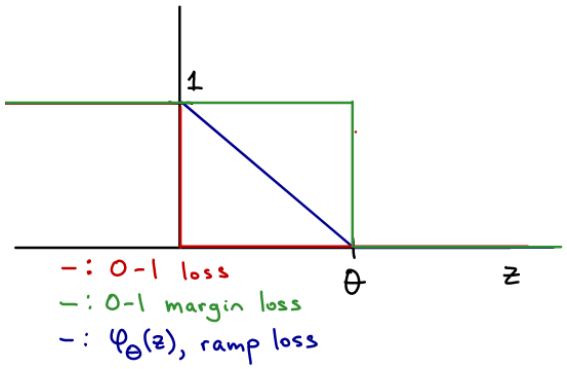

Figure 2: Graph showing showing the three three loss functions: $0-1$ loss, margin $0-1$ loss, and the ramp loss given by $\varphi_\theta(z)$.

We can show the following:

1. With probability $1-\delta$, for every w such that $\|w\|_1 \leq B_1$

$$\mathbb{E}_{\mathcal{D}}\ell_{\theta,w}(x,y) \leq \mathbb{E}_S\ell_{\theta,w}(x,y) + \sqrt{\frac{\ln 2/\delta}{2m}} + 2\text{Rad}_n(\mathcal{F})$$

where

$$\mathcal{F} = \{\ell_{\theta,w} \mid \|w\|_1 \leq B_1\}$$

Note that we saw the result like this in the proof of the uniform convergence theorem, so we will skip its proof. The key points in the argument are:

- McDiaramid's inequality
- Symmetrization
- Introducing Rademacher random variables to get Rademacher complexity

You may see such argument again in HW2, problem 3.

2.

$$\mathbb{E}_{\mathcal{D}}\ell_{\theta,w}(x,y) \geq \mathbb{P}_{\mathcal{D}}(y\langle w,x\rangle \leq 0)$$
$$\mathbb{E}_S\ell_{\theta,w}(x,y) \geq \mathbb{P}_S(y\langle w,x\rangle \leq \theta)$$

Note that the first inequality uses the $0-1$ loss on the RHS, while the second inequality uses the margin $0-1$ loss.

3. Next class, we will focus on bounding

$$\text{Rad}_n(\mathcal{F}) \leq O\left(\frac{B_1 R_\infty}{\theta}\sqrt{\frac{\ln d/\delta}{m}}\right)$$

For this task, we will use the *contraction inequality* of Rademacher complexity.

The theorem follows by combining the above three items. □