CS 7545: Machine Learning Theory **Fall 2019**

Lecture 26: Margin Theory Sketch

Lecturer: Jacob Abernethy *Scribes: Jing Yu*

Disclaimer: *These notes have not been subjected to the usual scrutiny reserved for formal publications.*

-No class Dec 2 (Mon).
-Office Hours: Tuesday 11 a.m. in Klaus 2134.
-Final Exam: Dec 11 (Wed) 2.40–5.30 p.m.

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

26.1 Spectrally-Normalized Margin Bounds for NNs

A linear classifier is a function

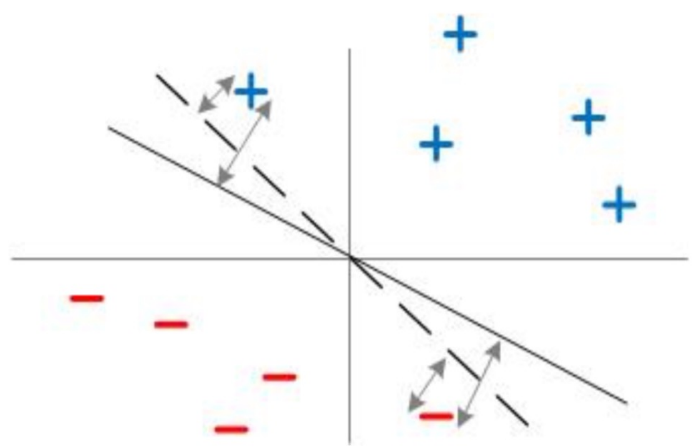
$$h_{\boldsymbol{w}}(\boldsymbol{x}) = \operatorname{sign}(\boldsymbol{w} \cdot \boldsymbol{x}).$$

which is parameterized by $\boldsymbol{w} \in \mathbb{R}^d$.

Given dataset $S = \{(\boldsymbol{x}_i, y_i) : i = 1, \ldots, m\}$ and classifier $h_{\boldsymbol{w}}$, the margin of $h_{\boldsymbol{w}}$ on S is

$$\rho(S) = \min_{i=1,\ldots,m} \frac{y_i(\boldsymbol{w} \cdot \boldsymbol{x}_i)}{\|\boldsymbol{w}\|_2}.$$

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.



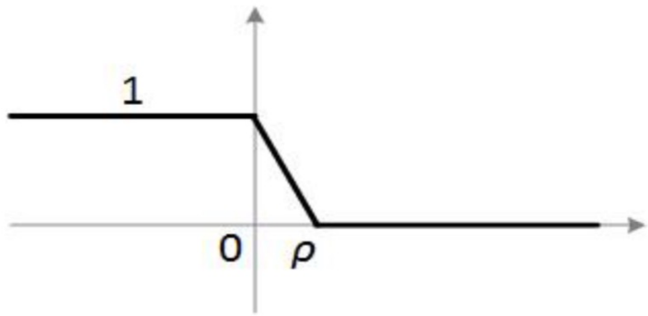
To find a tighter lower bound, consider the following class of linear classifiers with bounded norm:

$$H_\Lambda := \{h_{\boldsymbol{w}} : \mathbb{R}^d \to \{-1, 1\} : \|\boldsymbol{w}\|_2 \leq \Lambda\}.$$

Let $l(\hat{y}, y) = \varphi_\rho(\hat{y}y)$.

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

which looks like:



Fact: φ_ρ is $\frac{1}{\rho}$-Lipschitz.
Assume: WLOG, $\rho = 1$, otherwise, we can just shift to parameterize Λ.
Assume WLOG, $\|\boldsymbol{x}_i\|_2 \le 1, \forall i$.
Q: How well do margin penalized.
–classifiers generalize?
A: We know that this all boils down to Rademacher complexity of the loss class $l \circ H_\Lambda$.

Lemma 26.1 (Talagrad) $\hat{\mathcal{R}}_S(l \circ H_\Lambda) \le c\hat{\mathcal{R}}_{S|x}(H_\Lambda)$ *when l is c-Lipschitz.*

Proof: Refer the textbook. ■
Claim: for any data set S, any Λ, we have

$$\hat{\mathcal{R}}_S(H_\Lambda) \le \sqrt{\frac{\Lambda^2}{m}}.$$

Proof: Recall that

$$
\begin{aligned}
\mathcal{R}_S(H_\Lambda) :&= \mathbb{E}_{\sigma_1,\ldots,\sigma_m,\text{Rademacher r.v.s}}\left[\frac{1}{m}\sup_{h_{\boldsymbol{w}}\in H_\Lambda}\sum_{i=1}^{m} h_{\boldsymbol{w}}(\boldsymbol{x}_i)\sigma_i\right] \\
&= \frac{1}{m}\,\mathbb{E}_{\sigma_{1:m}}\left[\sup_{\boldsymbol{w}:\|\boldsymbol{w}\|\le\Lambda} \boldsymbol{w}(\sum_i \boldsymbol{x}_i\sigma_i)\right] \\
&\le \frac{1}{m}\,\mathbb{E}_{\sigma_{1:m}}\left[\sup_{\boldsymbol{w}:\|\boldsymbol{w}\|\le\Lambda} \|\boldsymbol{w}\|_2 \left\|\sum_i \boldsymbol{x}_i\sigma_i\right\|_2\right] \\
&= \frac{\Lambda}{m}\,\mathbb{E}_{\sigma_{1:m}}\left[\sqrt{\left\|\sum_i \sigma_i\boldsymbol{x}_i\right\|_2^2}\right] \\
&\le \frac{\Lambda}{m}\sqrt{\mathbb{E}_{\sigma_{1:m}}\left[\left\|\sum_i \boldsymbol{x}_i\sigma_i\right\|_2^2\right]} \\
&= \frac{\Lambda}{m}\sqrt{\mathbb{E}_{\sigma_{1:m}}\left[\sum_{i,j} \sigma_i\sigma_j\boldsymbol{x}_i\boldsymbol{x}_j\right]} \\
&= \frac{\Lambda}{m}\sqrt{\mathbb{E}_{\sigma_{1:m}}\left[\sum_i \sigma_i^2\,\|\boldsymbol{x}_i\|_2^2\right]} \\
&\le \frac{\Lambda}{\sqrt{m}}
\end{aligned}
$$

■