EECS 598: Statistical Learning Theory, Winter 2014 **Topic 16**

Margin Bounds

Lecturer: Clayton Scott *Scribe: Robert Vandermeulen*

1 Introduction

In these notes we will establish margin bounds. These are bounds that guarantee good generalization error (small $R(f)$) for any classifier f that correctly classifies the training data with a large confidence, where confidence is measured in terms of the functional margin $yf(x)$. We will also apply margin bounds to kernel-based classifiers, in which case they provide a sufficient condition for good generalization that is independent of the dimension of the feature space.

2 Margin Losses

Definition 1. *Define*

$$\phi_\rho(u) = \begin{cases} 1, & u \in (-\infty, 0) \\ 1 - \frac{u}{\rho}, & u \in [0, \rho] \\ 0, & u \in (\rho, \infty) \end{cases}.$$

The ρ-margin loss *is*

$$L_\rho(y, t) = \phi_\rho(y \cdot t).$$

Notice that

$$\begin{aligned} \widehat{R}_{L_\rho}(f) &= \frac{1}{n} \sum_{i=1}^n \phi_\rho(y_i f(x_i)) \\ &\leq \frac{1}{n} \sum_{i=1}^n \mathbb{1}\{y_i f(x_i) \leq \rho\} \\ &=: \widehat{R}_\rho(f). \end{aligned}$$

Here $\widehat{R}_\rho(f)$ represents the fraction of the training points misclassified or correctly classified with a "confidence" less than or equal to ρ. See Fig. 1.

Note. If $f(x) = w^T x$ and $\|w\|_2 = 1$, then $|w^T x|$ is the distance from x to the hyperplane defined by $w = \{x' : x'^T w = 0\}$, which is known as the "geometric margin." So when $\|w\|_2 = 1$, the functional and geometric margins coincide. More generally, they are proportional.

The basic margin bound is as follows.

Theorem 1. *Let $\mathcal{F} \subseteq [a, b]^{\mathcal{X}}$ and fix $\rho > 0$, $\delta > 0$. With probability at least $1 - \delta$, for all $f \in \mathcal{F}$*

$$R(f) \leq \widehat{R}_\rho(f) + \frac{2}{\rho} \mathcal{R}_n(\mathcal{F}) + \sqrt{\frac{\log \frac{1}{\delta}}{2n}}.$$

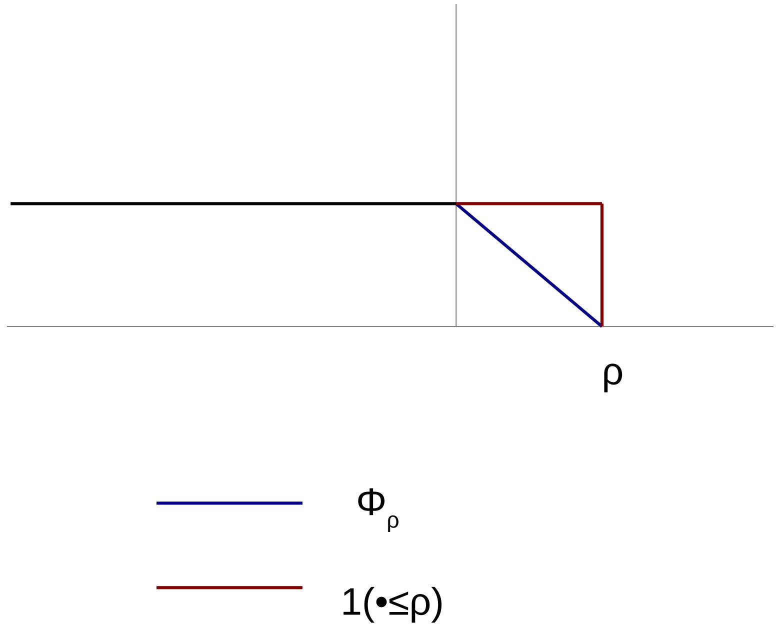

Figure 1: Losses

Proof. The proof follows from the one-sided Rademacher complexity bound applied to the class $\mathcal{G}$ of functions of the form $g(Z) = L_\rho(Y, f(X))$ where $f \in \mathcal{F}$, together with the Lipschitz composition property of Rademacher complexity, and the observations that $R(f) \leq R_{L_\rho}(f)$, $L_\rho(y, \cdot)$ has Lipschitz constant $\frac{1}{\rho}$, and $\mathcal{G} \subset [0,1]^{\mathcal{X}}$. $\square$

The interpretation of this result is that if $\widehat{R}_\rho(f)$ is small for large ρ then $R(f)$ is small. If $\widehat{R}_\rho(f)$ is small for large ρ, then we say f *has a large margin.*

One drawback of the above result is that it assumes ρ is fixed. This is not ideal, because the optimal ρ is not known a priori. However, with a little more work, it is also possible to obtain a margin bound that holds uniformly for all $\rho > 0$.

Theorem 2. *Let $\mathcal{F} \subseteq [-c, c]^{\mathcal{X}}$. With probability at least $1 - \delta$, for all $f \in \mathcal{F}$ and for all $\rho > 0$,*

$$R(f) \leq \widehat{R}_\rho(f) + \frac{4}{\rho}\mathcal{R}_n(\mathcal{F}) + \sqrt{\frac{\log\log_2(\frac{2c}{\rho})}{n}} + \sqrt{\frac{\log(\frac{2}{\delta})}{n}}$$

Proof. First note that if $\rho > c$, then $\widehat{R}_\rho(f) = 1$ in which case the bound holds trivially. So consider $0 < \rho \leq c$.

Let $(\rho_k)_{k\geq 1}$ and $(\varepsilon_k)_{k\geq 1}$ be positive sequences. Select $\varepsilon_k = \varepsilon + \sqrt{\frac{\log k}{n}}$. By Theorem 1 and the union bound, we have

$$
\begin{aligned}
P\left(\exists k \geq 1,\ f \in \mathcal{F} : R(f) - \widehat{R}_{\rho_k}(f) > \frac{2}{\rho_k}\mathcal{R}_n(\mathcal{F}) + \varepsilon_k\right) &\leq \sum_{k=1}^{\infty} \exp(-2n\varepsilon_k^2) \\
&= \sum_{k=1}^{\infty} \exp\left(-2n\left(\varepsilon + \sqrt{\frac{\log k}{n}}\right)^2\right) \\
&\leq \sum_{k=1}^{\infty} \exp\left(-2n\varepsilon^2\right)\exp\left(-2\log k\right) \\
&= \exp\left(-2n\varepsilon^2\right)\sum_{k=1}^{\infty} \exp\left(-2\log k\right) \\
&= \exp\left(-2n\varepsilon^2\right)\sum_{k=1}^{\infty} \frac{1}{k^2} \\
&= \frac{\pi^2}{6}\exp\left(-2n\varepsilon^2\right) \\
&\leq 2\exp\left(-2n\varepsilon^2\right).
\end{aligned}
$$

Let $\rho_k = c2^{-k}$. Then for all $0 < \rho \leq c$, there exists $k = k(\rho)$ such that $\rho \in (\rho_k, \rho_{k-1}]$. For $k = k(\rho)$ note that $\rho \leq \rho_{k-1} = 2\rho_k$ and so $\frac{1}{\rho_k} \leq \frac{2}{\rho}$ and $\sqrt{\log k} = \sqrt{\log\log_2\left(\frac{c}{\rho_k}\right)} \leq \sqrt{\log\log_2\left(\frac{2c}{\rho}\right)}$. Also for $k = k(\rho)$, $\widehat{R}_{\rho_{k(\rho)}}(f) \leq \widehat{R}_\rho(f)$. Putting these things together, with probability at least $1 - 2\exp\left(-2n\varepsilon^2\right)$ we have for all $\rho > 0$ and $f \in \mathcal{F}$

$$R(f) \leq \widehat{R}_\rho(f) + \frac{4}{\rho}\mathcal{R}_n(\mathcal{F}) + \sqrt{\frac{\log\log_2(\frac{2c}{\rho})}{n}} + \varepsilon.$$

To finish the proof take $\varepsilon = \sqrt{\frac{\log\frac{2}{\delta}}{2n}}$. $\square$

3 Application to Kernel Classes

Let's apply the preceding theory in the case where $\mathcal{F}$ is a ball in a reproducing kernel Hilbert space.

Corollary 1. *Let k be a kernel on $\mathcal{X}$ with $\sup_{x\in\mathcal{X}}\sqrt{k(x,x)} = B < \infty$. Let $M > 0$ be fixed. Then for all $\delta > 0$, with probabilitiy at least $1 - \delta$, for all $f \in B_k(M)$ and $\rho > 0$*

$$R(f) \leq \widehat{R}_\rho(f) + \frac{4}{\rho}\frac{BM}{\sqrt{n}} + \sqrt{\frac{\log\log_2\frac{2MB}{\rho}}{n}} + \sqrt{\frac{\log\frac{2}{\delta}}{2n}}.$$

Proof. Observe

$$|f(x)| \leq |\langle f, k(x,\cdot)\rangle| \leq \|f\|\|k(x,\cdot)\| \leq MB$$

and apply the previous theorem together with the Rademacher complexity bound for balls in an RKHS from the previous lecture. $\square$

We can specialize this result to linear classifiers as follows.

Corollary 2. *Let* $\mathcal{X} = \left\{x \in \mathbb{R}^d : \|x\|_2 \leq B\right\}$ *and* $\mathcal{F} = \{x \mapsto \langle x, w\rangle : \|w\|_2 \leq M\}$ *(here the inner product is just the dot product). Then with probability at least* $1-\delta$, *for all* $f \in \mathcal{F}$ *and* $\rho > 0$,

$$R(f) \leq \widehat{R}_\rho(f) + \frac{4}{\rho}\frac{BM}{\sqrt{n}} + \sqrt{\frac{\log\log_2 \frac{2BM}{\rho}}{n}} + \sqrt{\frac{\log\frac{2}{\delta}}{2n}}.$$

Proof. Note that the identity map $Id : \mathcal{X} \to \mathbb{R}^d$ is a valid feature map for the dot product kernel. Furthermore, the mapping $w \mapsto (x \mapsto \langle x, w\rangle)$ is injective. From our discussion of reproducing kernel Hilbert spaces (Topic 12), $\mathcal{F}$ is a ball in the reproducing kernel Hilbert space associated with the dot product kernel, and $\|w\|_2$ is the RKHS norm associated with the function $x \mapsto \langle x, w\rangle$. Now apply the previous corollary. $\square$

This bound is interesting because it gives a sufficient condition (small $\widehat{R}_\rho$ for large ρ) that guarantees good generalization that is independent of the dimension d. This contrasts with a VC bound, in which the right-hand side depends on the VC dimension of the set of linear classifiers (passing through the origin), which is d.

Also note that we can further specialize the previous result, focusing only on linear classifiers with $\|w\|_2 = M = 1$. Then $\widehat{R}_\rho(f)$ is the fraction of training points that are either misclassified, or that are within a distance of ρ to the hyperplane.