III. Guide to lectures for third two weeks.

Updated February 21, 2017

The focus of this two week period is Chapter 6 of the course notes. It continues development of the performance of empirical risk minimization for the specific context of binary classification. Function learning is a topic of the next chapter.

Tuesday, 2/14/17. The lecture was loosely based on Sections 6.1-6.3. The main point is that the ERM algorithm for the set of classifiers based on thresholding a set of m functions $\psi_1, \ldots, \psi_m$) (that is, a Dudley class of concepts) is PAC. However, it has two problems. It is not a very rich class of classifiers, and computation of the ERM hypothesis (i.e. minimizing the empirical risk) is NP hard, at least for the family of thresholded linear functions. The introduction of surrogate loss functions can address both problems. The lecture began with a summary of the fundamental theorem of concept learning (see part II of these notes from the preceding two weeks), and discussed how a different sort of guarantee, analogous to confidence intervals in statistics, follows from the same machinery with a different version of the mismatched minimization lemma (see Section 1). Then, two different sets of classifiers are discussed that are based on computing a real valued function from the features and then thresholding – see Sections 2 and 3 below.

No lecture Thursday, 2/16/17 and no recitation session 2/17/17 due to CSL Student Conference.

Tuesday, 2/23/17 This lecture is summarized in Section 4 (covering Section 6.1-6.3 of the notes) about classification with a surrogate loss function. However, we will begin the lecture by discussing two tools that are involved: (1) the contraction principle for Rademacher averages and its proof (In Section 5 of https://courses.engr.illinois.edu/ece543/supplement2.pdf, and (2) both the single version and double version of the performance bound of estimators in the abstract framework, described in Corollary 1 below. In Section 4, we first formulate a general theorem about the performance of classifiers using surrogate loss, and then examine two applications: concept classes formed by combining simpler classifiers, and concept classes arising from kernals.

Thursday, 2/23/17 Complete Section 4 if necessary, and briefly discuss problem set 3. Then the lecture discusses kernal methods following Section 6.4 of the course notes. These are used for classification problems as described at the end of Section 4 below as well as for regression with quadratic cost, as described in Section 7 of the course notes, which we will see next week.

1 Data based confidence bounds based on single version of mismatched minimization lemma

Suppose we'd like to find a minimizer of a function G defined on some domain $\mathcal{U}$, but the function G is not known. The following is an expanded version of the lemma from problem set one.

Lemma 1. *(Mismatched minimization lemma) Suppose that $\widehat{G}$ is an ϵ uniform approximation of G for some $\epsilon > 0$, meaning that $|G(u) - \widehat{G}(u)| \leq \epsilon$ for all $u \in \mathcal{U}$.*
(a) (Single version) For any $\widehat{u} \in \mathcal{U}$, $G(\widehat{u}) \leq \widehat{G}(\widehat{u}) + \epsilon$.
(b) (Double version) Suppose that u^* is a minimizer of $\widehat{G}$, meaning that $u^* \in \mathcal{U}$ and $\widehat{G}(u^*) \leq \widehat{G}(u)$ for all $u \in \mathcal{U}$. Then $G(u^*) \leq \inf_{u \in \mathcal{U}} G(u) + 2\epsilon$.

Proof. Part (a) is immediate from the assumption $|G(u) - \widehat{G}(u)| \leq \epsilon$ for all $u \in \mathcal{U}$. For any $u \in \mathcal{U}$, $G(u) \geq \widehat{G}(u) - \epsilon \geq \widehat{G}(u^*) - \epsilon \geq G(u^*) - 2\epsilon$. Therefore, $\inf_{u \in \mathcal{U}} G(u) \geq G(u^*) - 2\epsilon$, which is equivalent to part (b). □

While part (a), the single version of the lemma, is rather obvious, it is also useful in applications. In applications the point $\widehat{u}$ could be a minimizer or approximate minimizer of $\widehat{G}$, where $\widehat{G}$ is an approximation

of G computed from data. An advantage of the upper bound in part (a) is that it is computable from the data.

First Illustration To illustrate, we can state a variation of the Fundamental Theorem of Concept Learning to give upper bounds computable from the data:

Theorem 1. *(Confidence bounds for concept learning) Consider an agnostic concept learning problem* $(\mathsf{X}, \mathcal{P}, \mathcal{C})$*, and let* $\delta > 0$*. For any* $P \in \mathcal{P}$*, the ERM algorithm satisfies*

$$L_P(\widehat{C}_n) \leq L_{\mathsf{P}_n}(\widehat{C}_n) + 4\sqrt{\frac{V(\mathcal{C})\log(n+1)}{n}} + \sqrt{\frac{\log\left(\frac{1}{\delta}\right)}{2n}} \tag{1}$$

with probability at least $1-\delta$*. There is a universal constant* C *so that for any probability distribution* P *on* Z *and* $\delta \in (0,1)$*, the ERM algorithm satisfies*

$$L_P(\widehat{C}_n) \leq L_{\mathsf{P}_n}(\widehat{C}_n) + C\sqrt{\frac{V(\mathcal{C})}{n}} + \sqrt{\frac{\log\left(\frac{1}{\delta}\right)}{2n}} \tag{2}$$

with probability at least $1-\delta$*.*

In contrast to the fundamental theorem for concept learning stated at the end of the supplement notes for the second two weeks, the right hand sides of the bounds (1) and (2) begin with the empirical risk $L_{\mathsf{P}_n}(\widehat{C}_n)$ of the estimator rather than the generalized minimum risk $L_P^*(\mathcal{C})$, and the other terms on the right hand sides are smaller by a factor of two because they only need to upper bound $\triangle_n(Z^n)$ instead of $2\triangle_n(Z^n)$. The meaning of the bounds in (1) and (2) is similar to the meaning of confidence intervals for statistics. The bounds mean that, before the data is generated, no matter which $P \in \mathcal{P}$ is true, we expect that the bounds will hold with probability at least $1-\delta$. It is not correct or meaningful to say, after the data has been collected and after $L_{\mathsf{P}_n}(\widehat{C}_n)$ has been computed, that the bounds are true with probability at least $1-\delta$.

Note that the fundamental theorem for concept learning implies PAC learnability of the concept class by ERM if $\mathcal{C}$ is a VC class. Theorem 1 does not imply PAC learnability – it is simply a different type of result. The proof of Theorem 1 follows exactly the same lines as the proof of the fundamental theorem of concept learning, except that the bound $L_P(\widehat{C}_n) \leq L_{\mathsf{P}_n}(\widehat{C}_n) + \triangle_n(Z^n)$ is used instead of $L_P(\widehat{C}_n) \leq L_P^*(\mathcal{C}) + 2\triangle_n(Z^n)$.

Second Illustration Let's give another example of the use of both versions of the mismatched minimization lemma. Recall that in the general abstract formulation of ERM algorithms, the learning problem is simply specified by $(\mathsf{Z}, \mathcal{P}, \mathcal{F})$. The minimum possible expected risk for a $P \in \mathcal{P}$ is denoted by $L_P^*(\mathcal{F}) = \inf_{f\in\mathcal{F}} P(f)$. The empirical risk minimization (ERM) algorithm in this setup is given by

$$\widehat{f}_n = \arg\min_{f\in\mathcal{F}} \mathsf{P}_{Z^n}(f) = \arg\min_{f\in\mathcal{F}} \frac{1}{n}\sum_{i=1}^n f(Z_i).$$

Using

$$\triangle_n(Z^n) = \sup_{f\in\mathcal{F}} |\mathsf{P}_{Z^n}(f) - P(f)| = \sup_{f\in\mathcal{F}} \left|\frac{1}{n}\sum_{i=1}^n f(Z_i) - P(f)\right|. \tag{3}$$

the two versions of the mismatched minimization lemma imply that

$$L_P(\widehat{f}_n) \leq \mathsf{P}_{Z^n}(\widehat{f}_n) + \triangle_n(Z^n) \text{ (single version)}$$
$$L_P(\widehat{f}_n) \leq L_P^*(\mathcal{F}) + 2\triangle_n(Z^n) \text{ (double version) .}$$

The theorem based on the Vapnik-Chernovenkis symmetrization trick states that for any $P \in \mathcal{P}$, $\mathbb{E}[\triangle_n(Z^n)] \leq 2E_{P^n}[R_n(\mathcal{F}(Z^n))]$. For a bit more generality than the first time around, lets suppose that each $f : \mathsf{X} \to [a,b]$ for some interval $[a,b]$ with width $B = b-a > 0$. Then applying the fact that $\triangle_n(Z^n)$

has the bounded differences property with constant $c = \frac{B}{n}$ and McDiarmid's inequality yields that, for any $P \in \mathcal{P}$,

$$\triangle_n(Z) \leq \mathbb{E}\left[\triangle_n(Z)\right] + B\sqrt{\frac{2\log\left(\frac{1}{\delta}\right)}{n}} \qquad \text{with prob. at least } 1-\delta.$$

We thus get the following corollary (in original version, only the double version was included).

Corollary 1. *Given an abstract ERM learning problem $(\mathsf{Z}, \mathcal{P}, \mathcal{F})$ such that each $f : \mathsf{Z} \to [a, b]$ with $B = b - a$ for each $f \in \mathcal{F}$, and any probability distribution P on Z and $\delta \in (0, 1)$, with probability at least $1 - \delta$, both of the following statements are true:*

$$P(\widehat{f}_n) \leq \mathsf{P}_{Z^n}(\widehat{f}_n) + 2E_{P^n}[R_n(\mathcal{F}(Z^n))] + B\sqrt{\frac{\log\left(\frac{1}{\delta}\right)}{2n}} \qquad \textit{(for any algorithm)}$$

$$P(\widehat{f}_n) \leq L_P^*(\mathcal{F}) + 4E_{P^n}[R_n(\mathcal{F}(Z^n))] + B\sqrt{\frac{2\log\left(\frac{1}{\delta}\right)}{n}} \qquad \textit{(if ERM algorithm is used)}$$

The first bound in Corollary 1 is a data based confidence upper bound that, with probability at least $1 - \delta$, holds uniformly for all choices of $\widehat{f}_n$. The second bound in Corollary 1 shows that ERM is PAC if the term $E_{P^n}[R_n(\mathcal{F}(Z^n))]$ can be shown to converge to zero.

2 Revisiting Dudley classifiers

Chapter 6 of the course notes focuses on $\{-1, 1\}$ valued classifiers, so instead of using the function $\mathsf{pos}(x) = \mathbf{1}_{\{x \geq 0\}}$ we shall use the function $\mathsf{sgn}(x) = \mathbf{1}_{\{x \geq 0\}} - \mathbf{1}_{\{x < 0\}}$. With this change, a *Dudley class* of concepts for X is a set of $\{-1, 1\}$ valued functions (equivalent to a set of subsets of X) of the form $f(x) = \mathsf{sgn}\left(\sum_{i=1}^m c_i\psi_i(x) + h(x)\right)$, for some fixed $m \geq 1$, and fixed set of real-valued functions $\psi_1, \ldots, \psi_m$, and h. The functions are parameterized by $(c_1, \ldots, c_m) \in \mathbb{R}^m$. The Dudley class can be expressed as $\mathsf{sgn}(\mathcal{G} + h)$, where $\mathcal{G}$ is the span of $\psi_1, \ldots, \psi_m$.

As shown in the notes and problem set 2 (for nonzero h), the VC dimension of the Dudley class $\mathsf{sgn}(\mathcal{G} + h)$ is equal to the linear dimension of $\mathcal{G}$, which is less than or equal to m.

Corollary 2. *Consider an agnostic concept learning problem $(\mathsf{X}, \mathcal{P}, \mathcal{C})$ such that $\mathcal{C}$ is a Dudley class with dimension m. Then the Fundamental Theorem of Concept Learning and Theorem 1, on confidence bounds for concept learning, hold with $V(\mathcal{C}) = m$.*

This corollary is nice, but has two serious shortcomings. First, the dimension of the Dudley class must be finite, which is restrictive (for example, compare to the classifiers in Section 3 below) and could therefore impose a large inductive bias. Secondly, computation of the ERM classifier is computationally difficult (NP hard). The use of surrogate loss functions discussed in Section 4 below addresses both these points with a large degree of success.

3 Combined (aka weighted voting) classifiers – revisiting conv$(\mathcal{G})$

This section points to a much richer set of classifiers than the Dudley class.

Let $\mathcal{G}$ be a set of base classifiers mapping X to the label set $\{-1, 1\}$. These could be rather simple, such as interval classifiers for $\mathsf{X} = \mathbb{R}$, or more generally, $\mathcal{G}$ could be a Dudley set of classifiers of finite dimension.

Let $\mathcal{G}_1$ be the set of classifiers of the form $g(x) = \mathsf{sgn}\left(\sum_{i=1}^N c_i g_i(x)\right)$, where $N \geq 1$, $g_i \in \mathcal{G}$ for $i \in [N]$, and $(c_1, \ldots, c_N)$ is a probability vector. That is, $\mathcal{G}_1 = \mathsf{sgn}(\text{conv}(\mathcal{G}))$. Thus, a g in $\mathcal{G}_1$ is the result of comparing a convex combination of arbitrarily many simple classifiers to the threshold 0.

Often the VC dimension, $V(\mathcal{G}_1)$, is infinite. (See problem set 3 for the case $\mathcal{G}$ is the set of interval classifiers on $\mathbb{R}$.)

An even richer set of classifiers is $\mathcal{G}_2 \triangleq \mathsf{sgn}(\text{absconv}(\mathcal{G}))$. Since $\mathsf{sgn}(\lambda x) \equiv \mathsf{sgn}(x)$ for any $\lambda > 0$, it follows that $\mathcal{G}_2 = \mathsf{sgn}(\text{linear span}(\mathcal{G}))$.

The sets of classifers $\mathcal{G}_1$ and $\mathcal{G}_2$ can be much richer than Dudley classes because the set of possible functions $\mathcal{G}$ used for linear combinations can be infinite.

4 Classification with surrogate loss functions

Both the Dudley classifiers and the combined classifiers produce a real value u which is then input to the sgn function, to produce an estimate of the label $y \in \{-1, 1\}$. Therefore, the 0-1 loss can be written as a function of y and u:

$$\begin{aligned}
\ell(y,u) &= \mathbf{1}_{\{y \neq \mathsf{sgn}u\}} \\
&= \mathbf{1}_{\{y=1, u<0\}} + \mathbf{1}_{\{y=-1, u \geq 0\}} \\
&\leq \mathbf{1}_{\{y=1, u\leq 0\}} + \mathbf{1}_{\{y=-1, u \geq 0\}} \\
&= \mathbf{1}_{\{yu \leq 0\}} \\
&= \mathbf{1}_{\{-yu \geq 0\}}
\end{aligned}$$

The idea of surrogate loss function is to replace $\mathbf{1}_{\{x \geq 0\}}$ by a continuous, often convex, function that dominates it. That is, suppose $\varphi : \mathbb{R} \to \mathbb{R}_+$ is such that

1. φ is continuous
2. φ is nondecreasing
3. $\varphi(x) \geq \mathbf{1}_{\{x \geq 0\}}$ for all $x \in \mathbb{R}$.

It seems appropriate to call φ a *penalty function*, similar to use of the term for constrained optimization problems. The surrogate loss function, or φ-loss function, corresponding to penalty function φ is defined by $\ell_\varphi(y,u) \triangleq \varphi(-yu)$. Note that the surrogate loss function is greater than or equal to the original loss function derived from 0-1 loss: $\ell(y,u) \leq \ell_\varphi(y,u)$ for all $(y,u) \in \{-1,1\} \times \mathbb{R}$.

Table 1 displays some popular examples of penalty functions, their Lipschitz constants, and their corresponding loss functions.

Table 1: Popular penalty functions and corresponding Lipschitz constants and surrogate loss functions

Name	Penalty function $\varphi(x)$	M_φ	loss function $\ell_\varphi(y,u)$
exponential	e^x	–	e^{-yu}
logit	$\log_2(1+\mathrm{e}^x)$	$\frac{1}{\ln 2}$	$\log_2(1+\mathrm{e}^{-yu})$
hinge	$(1+x)_+$	1	$(1-yu)_+$
ramp	$\min\left\{1, \left(1+\frac{x}{\gamma}\right)_+\right\}$	$\frac{1}{\gamma}$	$\min\left\{1, \left(1-\frac{yu}{\gamma}\right)_+\right\}$

We will apply the contraction principle for Rademacher averages and Corollary 1 above (bounds for ERM based on Rademacher averages) to give performance bounds on ERM for concept learning with surrogate loss functions.

The learning problem is specified by $(\mathsf{X}, \mathsf{Y} = \{0,1\}, \mathcal{P}, \mathcal{F})$ where $\mathcal{F}$ is a family of functions mapping X to $\mathbb{R}$, and the classifiers having the form $\mathsf{sgn}(f)$ for $f \in \mathcal{F}$ for the original $0-1$ loss. We fix a penalty function φ and consider the surrogate loss function $\ell_\varphi(y,u) \triangleq \varphi(-yu)$. It turns out that this form of loss function has a nice property for bounding Rademacher averages, and as mentioned above, the surrogate loss is an upper bound on the 0-1 loss.

The result we derive is true for each $P \in \mathcal{P}$, so for the remainder of this section let P be a fixed probability distribution on $\mathsf{Z} = \mathsf{X} \times \{0,1\}$. We will not include subscript P in the notation. In order to avoid overuse of the letter "L," we write the φ-risk, or expected φ-loss, of $f \in \mathcal{F}$ as

$$A_\varphi(f) \triangleq \mathbb{E}\left[\varphi(-Yf(X))\right]$$

and its empirical version

$$A_\varphi(f) \triangleq \frac{1}{n}\sum_{i=1}^{n} \varphi(-Y_i f(X_i)).$$

The minimum φ-loss is denoted by

$$A^*_\varphi(\mathcal{F}) \triangleq \inf_{f\in\mathcal{F}} A_\varphi(f).$$

An algorithm for the problem is given by $\mathcal{A} = (A_n)_{n\geq 1}$ such that $A_n(X^n) = \widehat{f}_n$. We will give two bounds, based on the single and double versions of the mismatched minimization lemma. For the single version, $\mathcal{A}$ can be arbitrary, but typically it would be either an ERM or approximate ERM algorithm for the surrogate loss. For the double version, $\mathcal{A}$ should be the ERM algorithm for the surrogate loss.

Theorem 2. *(Bounds for classification using surrogate loss) Suppose $\mathcal{F}$ and the penalty function φ are chosen so that the following conditions are satisfied:*

1. *$\varphi(-yf(x)) \leq B$ for some constant B, for all $(x,y) \in \mathsf{X}\times\{0,1\}$ and $f \in \mathcal{F}$.*
2. *φ is M_φ-Lipschitz continuous for some constant M_φ.*

(Single version) For any n and $\delta \in (0,1)$, and any learning algorithm, the following bound holds with probability at least $1-\delta$:

$$L(\widehat{f}_n) \leq A_\varphi(\widehat{f}_n) \leq A_{\varphi,n}(\widehat{f}_n) + 4M_\varphi ER_n(\mathcal{F}(X^n)) + B\sqrt{\frac{\log(1/\delta)}{2n}} \tag{4}$$

(Double version) For any n and $\delta \in (0,1)$, and for the ERM algorithm for surrogate loss, the following bound holds with probability at least $1-\delta$:

$$L(\widehat{f}_n) \leq A_\varphi(\widehat{f}_n) \leq A^*_\varphi(\mathcal{F}) + 8M_\varphi ER_n(\mathcal{F}(X^n)) + B\sqrt{\frac{2\log(1/\delta)}{n}} \tag{5}$$

Proof. The fact $L(\widehat{f}_n) \leq A_\varphi(\widehat{f}_n)$ follows immediately from the fact the surrogate loss is greater than or equal to the 0-1 loss, so the remainder of the proof concentrates on the other two inequalities to be proved. We will cast the learning problem with the surrogate loss function into the abstract framework for ERM and apply Corollary 1. The appropriate class of functions on $\mathsf{Z} = \mathsf{X}\times\{0,1\}$ is the set of functions $\mathcal{H}_\varphi$ of the form $\ell_{\varphi,f}(x,y) = \varphi(-yf(x))$, for $f \in \mathcal{F}$. (That is, $\ell_{\varphi,f}(z)$ represents the surrogate loss for classifier f on a sample $(x,y) = z$.)

We could apply Corollary 1 immediately if we had a suitable upper bound on $ER_n(\mathcal{H}_\varphi)$. As an intermediate step, consider the class of functions $\mathcal{H}$ of the form $h(x,y) = -yf(x)$. We shall now show that the multiplicative structure of the loss function with $y \in \{-1,1\}$ implies that for any sample $Z^n = (X^n, Y^n)$, $R_n(\mathcal{H}(Z^n)) = R_n(\mathcal{F}(X^n))$. That is, given X^n, no matter how the n points $X_1,\ldots,X_n$ are labeled with ± 1's to get Y^n, the Rademacher average of $R_n(\mathcal{H}(Z^n))$ is the same. It is because if σ_i is a Rademacher random variable

independent of Y_i, then $\sigma_i Y_i$ has the same distribution as σ_i. In detail:

$$\begin{aligned} R_n(\mathcal{H}(Z^n)) &= \frac{1}{n} E_\sigma \left[\sup_{h \in \mathcal{H}} \left| \sum_{i=1}^n \sigma_i h(Z_i) \right| \right] \\ &= \frac{1}{n} E_\sigma \left[\sup_{h \in \mathcal{H}} \left| \sum_{i=1}^n \sigma_i Y_i f(X_i) \right| \right] \\ &= \frac{1}{n} E_\sigma \left[\sup_{h \in \mathcal{H}} \left| \sum_{i=1}^n \sigma_i f(X_i) \right| \right] \\ &= R_n(\mathcal{F}(X^n)) \end{aligned} \tag{6}$$

The next step is to apply the contraction principle for Rademacher averages. (See the supplementary notes for second two weeks.) A minor inconvenience is that $\varphi(0) \neq 0$. To work around that fact, we apply the contraction principle using the mapping $F : \mathbb{R}^n \to \mathbb{R}^n$ defined by $F(v)_i = \varphi(v_i) - \varphi(0)$. Then the contraction principle and the fact (6) imply:

$$R_n(F \circ \mathcal{H}(Z^n)) \leq 2M_\varphi R_n(\mathcal{H}(Z^n)) = 2M_\varphi R_n(\mathcal{F}(X^n)).$$

Note that $F \circ \mathcal{H}$ is not quite equal to $\mathcal{H}_\varphi$, but we do have $F \circ \mathcal{H} = \mathcal{H}_\varphi - \varphi(0)$. Application of the single version of Corollary 1 to the abstract learning problem $(\mathsf{Z}, \mathcal{P}, F \circ \mathcal{H})$ then yields

$$L(\widehat{f}_n) - \varphi(0) \leq A_\varphi(\widehat{f}_n) - \varphi(0) \leq A_{\varphi,n}(\widehat{f}_n) - \varphi(0) + 4M_\varphi E R_n(\mathcal{F}(X^n)) + B\sqrt{\frac{\log(1/\delta)}{2n}}$$

The terms $\varphi(0)$ on each side cancel out to yield (12). The bound (13) similarly follows from the double version of Corollary 1. □

4.1 Application of surrogate loss bounds (Theorem 2) for classifiers built from simple classifiers by voting

As an application, we can apply Theorem 2 to the class of combined (aka weighted voting) classifiers $\mathrm{conv}(\mathcal{G})$, or more generally $\mathrm{absconv}(\mathcal{G})$, described in Section 3. By basic properties of Rademacher averages we have $R_n(\lambda \mathrm{conv}\mathcal{G}(X^n)) = R_n(\lambda \mathrm{absconv}\mathcal{G}(X^n)) = \lambda R_n(\mathcal{G}(X^n))$ for all n with probability one. In turn, the Dudley's chaining technique (replacement of finite class lemma plus Sauer-Shelah lemma) implies there is an absolute constant C such that $\mathbb{E}\left[R_n(\mathrm{absconv}\mathcal{G}(X^n))\right] \leq C\sqrt{\frac{V(\mathcal{G})}{n}}$. Combining with Theorem 2 yields the following (where factors 4 and 8 are absorbed into the constant C):

Corollary 3. *Suppose $\mathcal{G}$ is a set of $\{1, -1\}$-valued classifiers for X and let $\mathcal{F} = \lambda conv(\mathcal{G})$ or $\mathcal{F} = \lambda absconv(\mathcal{G})$, where $\lambda > 0$. Let φ be a penalty function such that over the interval $[-\lambda, \lambda]$, $\mathbf{1}_{\{x \geq 0\}} \leq \varphi(x) \leq B$ and φ is M_φ-Lipschitz continuous.*
(Single version) For any n and $\delta \in (0, 1)$, and any learning algorithm, the following bound holds with probability at least $1 - \delta$:

$$L(\widehat{f}_n) \leq A_\varphi(\widehat{f}_n) \leq A_{\varphi,n}(\widehat{f}_n) + \lambda M_\varphi C\sqrt{\frac{V(\mathcal{G})}{n}} + B\sqrt{\frac{\log(1/\delta)}{2n}} \tag{7}$$

(Double version) For any n and $\delta \in (0, 1)$, and for the ERM algorithm for surrogate loss, the following bound holds with probability at least $1 - \delta$:

$$L(\widehat{f}_n) \leq A_\varphi(\widehat{f}_n) \leq A_\varphi^*(\mathcal{F}) + \lambda M_\varphi C\sqrt{\frac{V(\mathcal{G})}{n}} + B\sqrt{\frac{2\log(1/\delta)}{n}} \tag{8}$$

Margin based risk bound Specializing further, if φ is the ramp penalty function with parameter $\gamma > 0$, then $M_\varphi = 1/\gamma$, and the empirical surrogate loss $A_{\varphi,n}(f)$ can be bounded above for any $f \in \mathcal{F}$ as follows:

$$A_{\varphi,n}(f) = \frac{1}{n}\sum_{i=1}^{n} \varphi(-Y_i f(X_i)) \leq L_n^\gamma(f)$$

where $L_n^\gamma(f)$ is the empirical *margin error* of f defined by:

$$L_n^\gamma(f) \triangleq \frac{1}{n}\sum_{i=1}^{n} \mathbf{1}_{\{Y_i f(X_i) < \gamma\}}.$$

Notice that:

- For any $\gamma > 0$ and $f \in \mathcal{F}$, $L_n(f) \leq L_n^\gamma(f)$ (i.e. empirical 0-1 loss is bounded above by the empirical margin error)
- The function $\gamma \to L_n^\gamma(f)$ is nondecreasing

Also, for any $f \in \mathcal{F}$,

$$\begin{aligned} L_n^\gamma(f) &= L_n(f) + \frac{1}{n}\sum_{i=1}^{n} \mathbf{1}_{\{0 < Y_i f(X_i) < \gamma\}} + \frac{1}{n}\sum_{i=1}^{n} \mathbf{1}_{\{Y_i = 1, f(X_i) = 0\}} \\ &\leq L_n(f) + \frac{1}{n}\sum_{i=1}^{n} \mathbf{1}_{\{0 \leq Y_i f(X_i) < \gamma\}}, \end{aligned} \tag{9}$$

where the first term on the right-hand side of (9) is the empirical 0-1 loss, and the other terms give the fraction of training samples that were classified correctly but with only a small *margin* (i.e. values of $Y_i f(X_i)$ less than γ.)

Let $L^{\gamma,*}(\mathcal{F}) = \inf_{f\in\mathcal{F}} E[L^\gamma(f)]$, which is the minimim generalized risk using penalty function $\mathbf{1}_{\{yf(x)<\gamma\}}$. Note that $A_\varphi^*(\mathcal{F}) \leq L^{\gamma,*}(\mathcal{F})$ because $\varphi(x) \leq \mathbf{1}_{\{yf(x)<\gamma\}}$ for all x. For the ramp penalty function, $M_\varphi = 1/\gamma$ and $B = 1$, so Corollary 3 yields:

Theorem 3. *(Margin-based risk bound for weighted linear combinations) Suppose $\mathcal{G}$ is a set of $\{1, -1\}$-valued classifiers for X and let $\mathcal{F} = \lambda conv(\mathcal{G})$ or $\mathcal{F} = \lambda absconv(\mathcal{G})$, where $\lambda > 0$. Let φ be the ramp penalty function with width parameter γ. (Single version) For any $\gamma > 0$,*

$$L(\widehat{f}_n) \leq L_n^\gamma(\widehat{f}_n) + \frac{C\lambda}{\gamma}\sqrt{\frac{V(\mathcal{G})}{n}} + \sqrt{\frac{\log(1/\delta)}{2n}} \tag{10}$$

holds for all choices of $\widehat{f}_n$ with probability at least $1 - \delta$.
(Double version) For any $\gamma > 0$, and a classifier $\widehat{f}_n$ that minimizes the empirical surrogate risk,

$$L(\widehat{f}_n) \leq L^{\gamma,*}(\mathcal{F}) + \frac{C\lambda}{\gamma}\sqrt{\frac{V(\mathcal{G})}{n}} + \sqrt{\frac{2\log(1/\delta)}{n}} \tag{11}$$

with probability at least $1 - \delta$.

Remark 1. *The first terms on the righthand sides of (10) and (11) increase with γ and the second terms decrease with γ. The bound (11) implies that if γ is large and the ERM classifer $\widehat{f}_n$ nevertheless has a small expected fraction of correctly classified samples with margin less than γ, then its generalization error for the original 0-1 loss will be small.*

4.2 Application of surrogate loss bounds (Theorem 2) for classifiers from an RKHS

Section 6.4 of the course notes describes spaces of functions based on reproducing kernal hilbert spaces (RKHS). In this section we briefly indicate how such spaces of functions fit in with Theorem 2 for classification with surrogate loss.

It is assumed that the feature set X is a closed subset of $\mathbb{R}^d$ for some $d \geq 1$. Given a *Mercer kernal* $K : \mathsf{X} \times \mathsf{X} \to \mathbb{R}$, the RKHS associated with K is a set of functions on X with an inner product forming a hilbert space $(\mathcal{H}_K, \langle \cdot, \cdot \rangle_K)$. Let $\mathcal{F}_\lambda$ denote the closed ball of some radius $\lambda > 0$ in $\mathcal{H}_K$. The following are three important properties:

- For any $f \in \mathcal{F}_\lambda$, $\|f\|_\infty \leq \lambda C_K$ where $C_K = \sup_{x \in \mathsf{X}} \sqrt{K(x,x)}$.
- Let $P \in \mathcal{P}(\mathsf{X})$. If X^n has distribution P^n and X has distribution P, then $\mathbb{E}\left[R_n(\mathcal{F}_\lambda(X^n))\right] \leq \lambda \sqrt{\frac{\mathbb{E}[K(X,X)]}{n}}$.
- For the hinge penalty function $\varphi(x) = (x+1)^+$, computation of the ERM for the surrogate loss function ℓ_φ can be expressed as a quadratically constrained convex program, efficiently solved by the support vector machine algorithm.

With the first two properties above we can state the following Corollary of Theorem 2. The hinge penalty function $\varphi(x) = (1+x)_+$ is used, which is Lipschitz continuous with $M_\varphi = 1$, and we can take $B = \lambda C_K + 1$.

Corollary 4. *(Performance bound for RKHS and hinge surrogate loss function) Suppose $\mathcal{F}_\lambda$ is the closed ball of radius $\lambda > 0$ in an RKHS of functions on a closed set $\mathsf{X} \subset \mathbb{R}^d$ with associated Mercer kernal K. Suppose $C_K = \sup_{x \in \mathsf{X}} \sqrt{K(x,x)} < \infty$. Let $\varphi(x) = (1+x)_+$.*
(Single version) For any n and $\delta \in (0,1)$, and any learning algorithm, the following bound holds with probability at least $1-\delta$:

$$L(\widehat{f}_n) \leq A_\varphi(\widehat{f}_n) \leq A_{\varphi,n}(\widehat{f}_n) + 4\lambda \sqrt{\frac{\mathbb{E}\left[K(X,X)\right]}{n}} + (\lambda C_K + 1)\sqrt{\frac{\log(1/\delta)}{2n}} \tag{12}$$

(Double version) For any n and $\delta \in (0,1)$, and for the ERM algorithm for surrogate loss, the following bound holds with probability at least $1-\delta$:

$$L(\widehat{f}_n) \leq A_\varphi(\widehat{f}_n) \leq A^*_\varphi(\mathcal{F}_\lambda) + 8\lambda \sqrt{\frac{\mathbb{E}\left[K(X,X)\right]}{n}} + (\lambda C_K + 1)\sqrt{\frac{2\log(1/\delta)}{n}} \tag{13}$$