Rademacher Generalization Bounds

Amir Globerson

One of the key questions of machine learning is that of generalization. Namely, if we learn a model from a given training set, how well will it perform on a new test set. In the introductory ML course you learned some approaches to quantifying generalization via the concepts of PAC learning and VC dimensions. Here we will introduce another very useful tool in this context: Rademacher bounds. These are very useful for analyzing generalization in a wide array of settings, including large margin classification, regression, matrix factorization and others.

1 Reminder: Empirical Risk Minimization

Generalization is closely related to the difference between empirical averages and expected values. Consider for example learning from a labeled sampled $(x_1, y_1), \ldots, (x_n, y_n)$ with a hypothesis class $\mathcal{H}$ where each $h \in \mathcal{H}$ is a function (or hypothesis) $y = h(x)$. Now assume that (x, y) pairs are generated IID from distribution $p(x, y)$, and that we have a loss function $\ell(h, x, y)$ telling us much we lose from predicting $h(x)$ when the true answer is y. If we had known $p(x, y)$, we would have chosen the function h that minimizes the expected loss. Namely:

$$h_p^* = \arg\min_{h \in \mathcal{H}} \mathbb{E}_p \left[\ell(h, X, Y)\right] \equiv e_p(h) \tag{1}$$

Sadly, we cannot do this, since we do not know p. But we can approximate it using an average over n IID samples from p, namely:

$$e_{S_n}(h) = \frac{1}{n} \sum_{i=1}^{n} \ell(h, X_i, Y_i) \tag{2}$$

we note that $e_{S_n}(h)$ is a random variable. In expectation it will equal $e_p(h)$ but generally they will of course be different. The ERM approach it to minimize $e_{S_n}(h)$ with respect to h. Denote the ERM hypothesis by:

$$h^E(S_n) = \arg\min_{h \in \mathcal{H}} e_{S_n}(h) \tag{3}$$

The idea is that when we have enough samples, the curve $e_{S_n}(h)$ (as a function of h) will be quite close to $e_p(h)$ and thus minimizing the first is a good approximation of minimizing the second. This can be made precise via the following proposition.

Proposition 1.1: *Suppose a training sample S_n satisfies*

$$\sup_{h \in \mathcal{H}} |e_{S_n}(h) - e_p(h)| \leq \epsilon \, . \tag{4}$$

Then the hypothesis $h^E(S_n)$ will have loss that is 2ϵ close to the best loss. Namely:

$$e_p(h^E(S_n)) - e_p(h^*) \leq 2\epsilon \tag{5}$$

The sample S_n is a random variable, so when we sample it, it may or may not satisfy the property in Proposition 1.1. What we would like is for this property to be satisfied with high probability, so that we have high probability for our ERM classifier to be good.

The type of result that we would want therefore is that for any $\epsilon, \delta > 0$ there will exist an $n(\epsilon, \delta)$ such that for all $n \geq n(\epsilon, \delta)$ property Equation 4 will hold with high probability. Namely:

$$\mathbb{P}\left[\sup_{h \in \mathcal{H}} |e_{S_n}(h) - e_p(h)| \leq \epsilon\right] \geq 1 - \delta \tag{6}$$

If we have this guarantee, then with probability at least $1 - \delta$ our ERM classifier will have loss 2ϵ close to the optimal one.

2 Rademacher Complexity

Rademacher bounds are a powerful technical tool for obtaining bounds like Equation 6 for a wide range of losses and hypothesis classes.

In what follows we consider a single random variable Z rather than (X, Y). You may think of Z as (X, Y) but that's not necessary. Assume Z is distributed according to some $p(z)$. And denote n IID samples of Z by $S = Z_1, \ldots, Z_n$.

Now, say we have a set of functions $\mathcal{F}$ of Z, and we are interested in calculating $\mathbb{E}[f(Z)]$ for all $f \in \mathcal{F}$. It makes sense to estimate this expected value via an average $\mathbb{E}\left[\frac{1}{n}\sum_i f(Z_i)\right]$. The average is itself a random variable of course, and in expectation it is equal to the true expectation.

$$\mathbb{E}[f(Z)] = \mathbb{E}\left[\frac{1}{n}\sum_i f(Z_i)\right] \tag{7}$$

But of course for a particular sample of $Z_1, \ldots, Z_n$ they will be different. What we want is to consider their different and see what the probability is that it is high. Thus, we want to consider the random variable:

$$\Delta = \sup_{f \in \mathcal{F}} \mathbb{E}[f(Z)] - \frac{1}{n}\sum_i f(Z_i) \tag{8}$$

Note we did not use an absolute value for the difference. This also makes sense since we are more worried about the generalization error being larger than the empirical one.

We would like Δ to be small. As a first step it will be useful to know what $\mathbb{E}[\Delta]$ is. Namely, what in expectation is the maximum discrepancy between the empirical averages and the means (recall the source of randomness here is the sample S. Intuitively, as the class $\mathcal{F}$ grows, we expect Δ to grow as it is going to be harder to guarantee that the difference is small for all functions.

Rademacher complexity quantifies the "size" of a function class $\mathcal{F}$ by checking how easy it is to fit it to random values. We begin with the definition:

Definition 2.1: *Denote $\sigma_1, \ldots, \sigma_n$ IID random variables, each with distribution uniform in $\{-1, +1\}$. Given a sample $S = Z_1, \ldots, Z_n$ the Rademacher complexity of a function class $\mathcal{F}$ is defined as:*

$$\mathcal{R}(\mathcal{F}, S) = \mathbb{E}_\sigma \left[\sup_{f \in \mathcal{F}} \frac{1}{n} \sum_i \sigma_i f(Z_i) \right] \tag{9}$$

Note that we don't take expectation with respect to Z_i here. So more formally we could have written this as expectation conditioned on S. The important thing to note is that $\mathcal{R}(\mathcal{F}, S)$ itself is a random variable, as it depends on S which is random.

Here are two interpretations of $\mathcal{R}$:

- Assume $\mathcal{F}$ is a function whose values are $\{-1, +1\}$. Then $\mathcal{R}$ reflects the following process. Sample σ_i to be a sequence of $\{-1, +1\}$ values. Now look for a function in your class that is best correlated with those. For example if $n = 5$ and you sampled $\{-1, -1, 1, 1, 1\}$ and you have a function $f \in \mathcal{F}$ that assigns values $f(z_1) = -1, f(z_2) = -1, \ldots, f(z_5) = 1$, then $\sup_{f \in \mathcal{F}} \sum_i \sigma_i f(z_i) = 5$. If this is true for all σ then $\mathcal{R}(\mathcal{F}, S) = 5$ which is the maximum possible value, indicating that our class $\mathcal{F}$ is highly expressive. As the other extreme, consider the case where f contains only the constant function $f(z) = 1$. Then we can see that $\mathcal{R}(\mathcal{F}, S) = 0$ for all S, indicating low expressiveness.

- In our case the function $f(x, y)$ will actually measure how well the classifier x predicts y (e.g., we can have a hinge loss $f(x, y) = [1 - yh(x)]_+$ where $y \in \{-1, +1\}$ and $h \in \mathcal{H}$ is some prediction function). Here we can interpret the Rademacher complexity as $\sigma_i = -1$ indicating the loss should be small for (x_i, y_i) and $\sigma_i = -1$ indicating it should be large. So the -1 points are a "training" set and $+1$ is a test set. The sup_h seeks a worst case where error on training is small but error on test is high. If the class $\mathcal{F}$ is rich enough to achieve such test-train gaps, that indicates generalization will be difficult.

- Since σ_i has probability 0.5 for each value, the most likely configuration of the σ_i is such that there's an equal number of $+1$ and -1. Thus, $\mathcal{R}$ is like taking many random splits of our sample S, and checking how different the averages of f are for these two parts. For each random split, we consider the f that maximizes the difference. If averages on these splits are quite different, this intuitively implies that the difference from the expected value should also be large.

2.1 Uniform Convergence and Rademacher Complexity

A key theorem about Rademacher complexity relates it to the difference between empirical averages and expectations of function in the class $\mathcal{F}$.

Theorem 2.2: *Given a function class $\mathcal{F}$, the following holds:*

$$\mathbb{E}\left[\sup_{f\in\mathcal{F}}\frac{1}{n}\sum_i f(Z_i) - \mathbb{E}\left[f(Z)\right]\right] \leq 2\mathbb{E}\left[\mathcal{R}(\mathcal{F},S)\right] \tag{10}$$

Proof: The proof uses a nice trick of introducing an additional IID sample $S' = (Z'_1,\ldots,Z'_n)$. For any i we can write $\mathbb{E}\left[f(Z'_i)\right] = \mathbb{E}\left[f(Z)\right]$ and therefore:

$$\mathbb{E}\left[\sup_{f\in\mathcal{F}}\frac{1}{n}\sum_i f(Z_i) - \mathbb{E}\left[f(Z)\right]\right] = \frac{1}{n}\mathbb{E}\left[\sup_{f\in\mathcal{F}}\sum_i f(Z_i) - \mathbb{E}\left[\sum_i f(Z'_i)\right]\right]$$

Via convexity of sup we have:

$$\leq \quad \frac{1}{n}\mathbb{E}\left[\sup_{f\in\mathcal{F}}\sum_i f(Z_i) - \sum_i f(Z'_i)\right] = \frac{1}{n}\mathbb{E}\left[\sup_{f\in\mathcal{F}}\sum_i \left(f(Z_i) - f(Z'_i)\right)\right]$$

To introduce σ_i variables, note the following. Denote $V_i = f(Z_i) - f(Z'_i)$. Then V_i has the same distribution as $-V_i$. Therefore, for any setting of $\sigma_1,\ldots,\sigma_n$ we have that the set of random variables $V_1,\ldots,V_n$ has the same distribution as $\sigma_1 V_1,\ldots,\sigma_n V_n$. We can therefore write the above as:

$$\begin{aligned} &= \quad \frac{1}{n}\mathbb{E}_{\sigma,S,S'}\left[\sup_{f\in\mathcal{F}}\sum_i \sigma_i\left(f(Z_i) - f(Z'_i)\right)\right] \\ &\leq \quad \frac{1}{n}\mathbb{E}_{\sigma,S,S'}\left[\sup_{f\in\mathcal{F}}\sum_i \sigma_i f(Z_i) + \sup_{f\in\mathcal{F}} -\sum_i \sigma_i f(Z'_i)\right] \end{aligned}$$

where the second inequality follows from $\|a+b\|_\infty \leq \|a\|_\infty + \|b\|_\infty$ (the triangle inequality for the ℓ_∞ norm). We now arrive at the desired result since $\sigma_1,\ldots,\sigma_n$ has the same distribution as $-\sigma_1,\ldots,-\sigma_n$, and therefore the above two terms both equal $\mathbb{E}\left[\mathcal{R}(\mathcal{F},S)\right]$. $\square$

This result tells us that empirical means and expectations are close to within $2\mathbb{E}\left[\mathcal{R}(\mathcal{F},S)\right]$. But we only know its true in expectation. Namely, there could be "bad" samples S such that the difference is larger than $2\mathbb{E}\left[\mathcal{R}(\mathcal{F},S)\right]$, but when taking expectation over S, the expected difference will be smaller. than $2\mathbb{E}\left[\mathcal{R}(\mathcal{F},S)\right]$. In fact, we can use Equation 10 to claim that such "bad" samples are not likely. To see this, we present present a simple argument based on Markov inequality, and then later give a tighter bound in Theorem 2.4.

Begin with the Markov based argument. In the terminology of Equation 8 we simply have:

$$\mathbb{E}\left[\Delta\right] \leq 2\mathbb{E}\left[\mathcal{R}(\mathcal{F},S)\right] \tag{11}$$

Now we want to argue that Δ cannot be too far from its expected value. For any $\delta > 0$ we have by Markov's inequality that:

$$\mathbb{P}\left[\Delta \geq \frac{2\mathbb{E}\left[\mathcal{R}(\mathcal{F}, S)\right]}{\delta}\right] \leq \delta \tag{12}$$

This says if we want to know with probability greater than $1 - \delta$ that Δ is greater than some number, then this number grows as $\frac{1}{\delta}$. Turns out we can have a much better guarantee. To prove it, we need the McDiarmid inequality, stated next

Theorem 2.3 : *[McDiarmid's Inequality] Consider n independent random variables $X_1, \ldots, X_n$ (not necessarily identically distributed). Assume $g(x_1, \ldots, x_n)$ is a function such that:*

$$|g(x_1, \ldots, x_n) - g(x_1, \ldots, x_{i-1}, x_i', x_{i+1}, \ldots, x_n)| \leq c \tag{13}$$

for some c, and all values of $x_1, \ldots, x_n, x_i', i$. Namely, changing one variable can change the value of g by at most c. Then it holds that for all $\delta > 0$:

$$\mathbb{P}\left[g(X_1, \ldots, X_n) - \mathbb{E}\left[g(X_1, \ldots, X_n)\right] \geq c\sqrt{\frac{n}{2}\ln\frac{2}{\delta}}\right] \leq \delta \tag{14}$$

The inequality states that if g satisfies the condition in Equation 13, then with probability at least $1 - \delta$ it will be $O(\sqrt{\ln \delta^{-1}})$ close to its expected value.

We can now apply this to our Δ random variable, to get the following improved bound (in terms of its dependence on δ).

Theorem 2.4: *Let $\mathcal{F}$ be a function class where $|f(z)| \leq \alpha$ for all z and $f \in \mathcal{F}$. Then:*

$$\mathbb{P}\left[\Delta \geq 2\mathbb{E}\left[\mathcal{R}(\mathcal{F}, S)\right] + \alpha\sqrt{\frac{2}{n}\ln\frac{2}{\delta}}\right] \leq \delta \tag{15}$$

Proof: Consider Δ as a function of $Z_1, \ldots, Z_n$. Recall:

$$\Delta(Z_1, \ldots, Z_n) = \sup_{f \in \mathcal{F}} \mathbb{E}\left[f(Z)\right] - \frac{1}{n}\sum_{i=1}^{n} f_i(Z_i) \tag{16}$$

The function Δ will correspond to g in the McDiarmid inequality. Now consider a sample $s = z_1, \ldots, z_n$ and $s' = z_1, \ldots, z_{i-1}, z_i', z_{i+1}, z_n$ Then (because difference of sups is smaller than sup of differences):

$$\Delta(s) - \Delta(s') \leq \sup_{f \in \mathcal{F}} \frac{1}{n}(f_i(z_i) - f_i(z_i')) \leq \frac{2\alpha}{n} \tag{17}$$

And similarly $\Delta(s') - \Delta(s) \leq \frac{2\alpha}{n}$. We conclude that Δ satisfies the McDiarmid condition with $c = \frac{2\alpha}{n}$. Therefore:

$$\begin{aligned}
\mathbb{P}\left[\Delta - \mathbb{E}\left[\Delta\right] \geq \frac{2\alpha}{n}\sqrt{\frac{n}{2}\ln\frac{2}{\delta}}\right] &\leq \delta \\
\mathbb{P}\left[\Delta - \mathbb{E}\left[\Delta\right] \geq \alpha\sqrt{\frac{2}{n}\ln\frac{2}{\delta}}\right] &\leq \delta
\end{aligned}$$

And now, since we know $\mathbb{E}[\Delta] \leq 2\mathbb{E}[\mathcal{R}(\mathcal{F}, S)]$ we conclude:

$$\mathbb{P}\left[\Delta - 2\mathbb{E}[\mathcal{R}(\mathcal{F}, S)] \geq \alpha\sqrt{\frac{2}{n}\ln\frac{2}{\delta}}\right] \leq \delta \tag{18}$$

□ Theorem 2.4 says that the difference Δ is close to $2\mathbb{E}[\mathcal{R}(\mathcal{F}, S)]$ with high probability, and the difference decreases with n, which we would expect.

2.2 Rademacher for Finite Classes

The VC dimension of a finite hypothesis class $\mathcal{H}$ is upper bounded by $\log|H|$. Turns out a similar result, known as Massart's lemma, holds for Rademacher complexity.

Theorem 2.5 : *Let $\mathcal{F}$ be a finite set of functions, and assume $|f(z)| \leq r$ for all z. Then for any S it holds that:*

$$\mathcal{R}(\mathcal{F}, S) \leq \sqrt{\frac{2r^2 \ln|\mathcal{F}|}{n}} \tag{19}$$

Proof can be found in most textbooks covering this topic (e.g., see [1]).

2.3 Rademacher for Bounded Norm Linear Classifiers

Consider the set of functions $f(x) = \boldsymbol{w} \cdot \boldsymbol{x}$ where $\|\boldsymbol{w}\|_2 \leq B$. Denote these by $\mathcal{F}$ The following theorem provides its Rademacher complexity.

Theorem 2.6 : *Let $\mathcal{F}$ be the class of linear functions, with bounded norm $\|\boldsymbol{w}\|_2 \leq B$. Assume all samples are such that $\|\boldsymbol{x}_i\|_2 \leq R$. Then:*

$$\mathcal{R}(\mathcal{F}, S) \leq \sqrt{\frac{R^2B^2}{n}} \tag{20}$$

Proof: We have that:

$$\begin{aligned}\mathcal{R}(\mathcal{F}, S) &= \mathbb{E}_\sigma\left[\sup_{\boldsymbol{w}:\|\boldsymbol{w}\|_2\leq B}\frac{1}{n}\sum_i \sigma_i \boldsymbol{w}\cdot\boldsymbol{x}_i\right] \\ &= \frac{1}{n}\mathbb{E}_\sigma\left[\sup_{\boldsymbol{w}:\|\boldsymbol{w}\|_2\leq B}\boldsymbol{w}\cdot\sum_i \sigma_i\boldsymbol{x}_i\right]\end{aligned}$$

This will be maximized when $\boldsymbol{w}$ is colinear with $\sum_i \sigma_i\boldsymbol{x}_i$ and has norm B. So we have:

$$\mathcal{R}(\mathcal{F}, S) \leq \frac{B}{n}\mathbb{E}\left[\|\sum_i \sigma_i\boldsymbol{x}_i\|_2\right] \leq \frac{B}{n}\mathbb{E}\left[\|\sum_i \sigma_i\boldsymbol{x}_i\|_2\right] = \frac{B}{n}\mathbb{E}\left[\sqrt{\|\sum_i \sigma_i\boldsymbol{x}_i\|_2^2}\right] \tag{21}$$

We now use the fact that $\sqrt{z}$ is a concave function and Jensen's inequality therefore implies $\mathbb{E}\left[\sqrt{Z}\right] \leq \sqrt{\mathbb{E}[Z]}$ for any random variable Z. we have:

$$\leq \quad \frac{B}{n}\sqrt{\mathbb{E}\left[\|\sum_i \sigma_i \boldsymbol{x}_i\|_2^2\right]} = \frac{B}{n}\sqrt{\mathbb{E}\left[\sum_{ij} \sigma_i\sigma_j \boldsymbol{x}_i \cdot \boldsymbol{x}_j\right]}$$

Now if $i \neq j$ then $\mathbb{E}[\sigma_i\sigma_j] = 0$ and if $i = j$ we have $\mathbb{E}[\sigma_i\sigma_i] = \mathbb{E}\left[\sigma_i^2\right] = 1$. Therefore:

$$\mathcal{R}(\mathcal{F}, S) \quad \leq \quad \frac{B}{n}\sqrt{\sum_i \|\boldsymbol{x}_i\|_2^2} \leq \frac{B}{n}\sqrt{nR^2} = \sqrt{\frac{R^2B^2}{n}}$$

concluding the proof. □

2.4 Rademacher Talagrand Contraction Lemma

For learning we are interested in the expected value of loss functions. Namely, given a hypothesis class $\mathcal{H}$ from inputs $\mathcal{X}$ to output $\mathcal{Y}$ and loss function $\ell : \mathcal{X} \to \mathcal{Y}$, we are interested in the functions $\ell(h(x), y)$. Thus, our function class of interest is:

$$\mathcal{F} = \{f(x, y) = \ell(h(x), y) : h \in \mathcal{H}\} \tag{22}$$

How do we calculate the Rademacher complexity of $\mathcal{F}$ from that of $\mathcal{H}$. The following result, known as the Talagrand contraction lemma, is very useful in this context.

Theorem 2.7 : *Let $\phi : \mathbb{R} \to \mathbb{R}$ be an L Lipschitz function. Namely: for all x, x' it holds that $|\phi(x) - \phi(x')| \leq |x - x'|$. Let $\mathcal{H}$ be a class of functions, and $\phi \circ \mathcal{H}$ be the composition of ϕ and $\mathcal{H}$. Then for all S*

$$\mathcal{R}(\phi \circ \mathcal{H}, S) \leq L\mathcal{R}(\mathcal{H}, S) \tag{23}$$

Next, we see how to apply this to the case of prediction losses. We focus on the hinge loss, but similar arguments can be made for log-loss (i.e., relative entropy) and others. Consider the hinge function:

$$\phi(v) = [1 - v]_+ \tag{24}$$

Then it's easy to see that it is Lipschitz with $L = 1$. Now, say we want to do classification on inputs $\boldsymbol{x}$. We consider a class of functions $\mathcal{H} : \boldsymbol{x} \to \mathbb{R}$ and our classifier for each $h \in \mathcal{H}$ is $y = \text{sign}[h(\boldsymbol{x})]$. Our goal is to find classifiers that minimize the prediction error. Namely, the expected error $\mathbb{E}[Y = \text{sign}[h(X)]]$. However, this is hard to optimize, so the hinge loss (or other convex surrogates) is often used.

$$\ell(h(\boldsymbol{x}), y) = \phi(yh(\boldsymbol{x})) \tag{25}$$

The corresponding function class is:

$$\mathcal{H}_\phi = \{f(x, y) = [1 - yh(x)]_+ : h \in \mathcal{H}\} \tag{26}$$

Proposition 2.8: *Let $\mathcal{H}_\phi$ be a function class as defined in Equation 26. Then*

$$\mathcal{R}(\mathcal{H}_\phi, S) \leq \mathcal{R}(\mathcal{H}, S) \tag{27}$$

Proof:

$$\mathcal{R}(\mathcal{H}_\phi, S) = \mathbb{E}_\sigma \left[\sup_{f \in \mathcal{H}_\phi} \frac{1}{n} \sum_i \sigma_i f(z_i) \right] = \mathbb{E}_\sigma \left[\sup_{h \in \mathcal{H}} \frac{1}{n} \sum_i \sigma_i \left[1 - y_i h(x_i) \right]_+ \right] \tag{28}$$

Applying the contraction lemma, we get:

$$\mathcal{R}(\mathcal{H}_\phi, S) \leq \mathbb{E}_\sigma \left[\sup_{h \in \mathcal{H}} \frac{1}{n} \sum_i \sigma_i y_i h(x_i) \right] = \mathbb{E}_\sigma \left[\sup_{h \in \mathcal{H}} \frac{1}{n} \sum_i \sigma_i h(x_i) \right] \tag{29}$$

Where in the last transition we used the fact that $y_i \in \{-1, +1\}$ is fixed here, and so the distribution of the random variable $(y_1\sigma_1, y_n\sigma_n)$ is identical to that of $\sigma_1, \ldots, \sigma_n$. We now conclude from definition that:

$$\mathcal{R}(\mathcal{H}_\phi, S) \leq \mathcal{R}(\mathcal{H}, S) \tag{30}$$

□

2.5 Generalization for Low Norm Linear Classifiers

The presentation here is similar to Chapter 26 in [1]. The SVM problem (in the hard SVM case) is the following optimization problem:

$$\begin{array}{ll} \min_{\boldsymbol{w}} & \|\boldsymbol{w}\|_2^2 \\ \text{s.t.} & y_i \boldsymbol{w} \cdot \boldsymbol{x}_i \geq 1 \ \ \forall i \end{array} \tag{31}$$

When presenting it, we motivated it by wanting to maximize the margin of the classifier, which we said intuitively should lead to better generalization. Here we would like to make this claim more formal, by showing that classifiers with lower norm do have better generalization guarantees.

We consider the following assumptions:

- All inputs will have $\|\boldsymbol{x}\| \leq R$ (under the true distribution $p(x, y)$, and therefore in training as well).
- There exists a classifier $\boldsymbol{w}^*$ which has zero generalization error. Namely $e_p(\boldsymbol{w}^*) \equiv \mathbb{E}\left[\text{sign}\left[\boldsymbol{w}^* \cdot X\right] = Y\right] = 0$.
- The norm of $\boldsymbol{w}^*$ is bounded by B.

We would like to show that the generalization error is a function only of B and not the dimension of $\boldsymbol{x}$.

First we apply the results of the previous sections to get a bound on the expected error of the hinge loss, in terms of the empirical hinge loss. Namely, we have that with probability greater than $1-\delta$ it holds for all $\boldsymbol{w}$ that:

$$\mathbb{E}\left[[1 - Y\boldsymbol{w}\cdot X]_+\right] \leq \frac{1}{n}\sum_i [1 - y_i \boldsymbol{w}\cdot \boldsymbol{x}_i]_+ + 2\sqrt{\frac{R^2B^2}{n}} + (1+RB)\sqrt{\frac{2}{n}\ln\frac{2}{\delta}} \quad (32)$$

To use this to express the generalization error of the SVM classifier we note:

- The norm of the SVM classifier will be smaller than B (since we know there is a norm B classifier that will separate the training data, and SVM minimizes the norm).
- Equation 32 holds for the SVM solution which we denote by $\boldsymbol{w}_{SVM}$.
- Denote the true classification error of the SVM classifier by $e_p(\boldsymbol{w}_{SVM})$. Then :

  $$e_p(\boldsymbol{w}_{SVM}) = \mathbb{E}\left[\text{sign}\left[\boldsymbol{w}_{SVM}\cdot\boldsymbol{x}\right] = Y\right] \leq \mathbb{E}\left[[1 - Y\boldsymbol{w}_{SVM}\cdot\boldsymbol{x}]_+\right] \quad (33)$$

  where the first equality is by definition, the second follows from the fact that hinge loss upper bounds the zero one loss.
- The empirical hinge loss for $\boldsymbol{w}_{SVM}$ is zero.

We conclude the following.

Proposition 2.9 : *The expected classification error of the SVM solution is upper bounded as follows:*

$$e_p(\boldsymbol{w}_{SVM}) \leq 2\sqrt{\frac{R^2B^2}{n}} + (1+RB)\sqrt{\frac{2}{n}\ln\frac{2}{\delta}} \quad (34)$$

This result means that if SVM is used in a case where a low norm solution exists, then its generalization error will be small. The key thing to note is that this bound does not depend on the dimension of $\boldsymbol{x}$ at all, and this justifies the use of kernels where this dimension may be infinite. Of course for SVM to generalize well, we still need there to be a low norm classifier in the kernel feature space, which may or may not be the case in practice.

References

[1] S. Shalev-Shwartz and S. Ben-David. *Understanding machine learning: From theory to algorithms*. Cambridge university press, 2014.