Rademacher complexity and the generalization bounds

Pannagadatta Shivaswamy and Tony Jebara

Abstract

For a modified version of RMM, we derive the empirical Rademacher complexities and then use them to derive the generalization bounds.

1 Introduction

Support Vector Machines [5] find hyperplanes of the form $\mathbf{w}^\top \mathbf{x} = 0$, $\mathbf{w} \in \mathbb{R}^m, \mathbf{x} \in \mathbb{R}^m$ as the decision boundary from a limited number of examples. We suppress the bias term in this addendum for simplicity. The decision boundary is found by minimizing a combination $\mathbf{w}^\top \mathbf{w}$ and an upper bound on the number of misclassifications.

Minimizing $\frac{1}{2}\mathbf{w}^\top \mathbf{w}$ can also be seen as choosing a function $g(\mathbf{x}) = \mathbf{w}^\top \mathbf{x} + b$ from a set of linear functions with bounded 2-norm. For a suitable choice of E, the SVM solution can be seen as choosing a function $g(\cdot)$ from the set $\{\mathbf{x} \rightarrow \mathbf{w}^\top \mathbf{x} | \frac{1}{2}\mathbf{w}^\top \mathbf{w} \leq E\}$. Rademacher complexity of a function class is a measure of how "simple" or "complex" a function class is. Further, it can be used to derive generalization bounds. For SVMs, such results can be found in [4].

In this addendum, for a slightly modified versions of RMM and $\mathbf{\Sigma}$-SVM, we define the function classes. For the function classes defined, we derive the empirical Rademacher complexity. Further, they will be used to show the generalization bounds. This addendum shows how the known bounds for SVM change with respect to RMM and $\mathbf{\Sigma}$-SVM.

Most of the material here closely follows the derivation of Rademacher complexity and the generalization bounds in [4].

2 The function classes

We assume that $(\mathbf{x}_i, y_i)_{i=1}^n$, with $\mathbf{x}_i \in \mathbb{R}^m$ and $y_i \in \{\pm 1\}$ is a training sample drawn independent and identically distributed (iid) from an unknown underlying distribution $\Pr[(\mathbf{x}, y)]$. A linear classifier (such as an SVM) estimates a function $g(\mathbf{x}) := \mathbf{w}^\top \mathbf{x}$ to match the sign of the labels to minimize the probability of error on future test data from $\Pr[(\mathbf{x}, y)]$.

SVMs maximize the margin by minimizing $\frac{1}{2}\mathbf{w}^\top \mathbf{w}$. This can be seen as choosing $g(\cdot)$ from a restricted class of functions via a 2-norm ball on $\mathbf{w}$. Given a choice of the parameter E in the SVM (where E plays the role of the regularization parameter), the set of linear functions that will be considered is:

Definition 1 $\mathcal{F}_E := \{\mathbf{x} \rightarrow \mathbf{w}^\top \mathbf{x} | \frac{1}{2}\mathbf{w}^\top \mathbf{w} \leq E\}$.

This set of functions ensures that the SVM recovers a certain absolute margin determined by E.

Let us now consider the function class for a modified version of RMM. Note that the proposed RMM bounds the projection on the training examples. However, the generalization bounds in this addendum hold only if we bound the projections on an independent set $\{\mathbf{u}_1, \mathbf{u}_2, \ldots \mathbf{u}_{n_u}\}$ – which can be from $\Pr[(\mathbf{x})]$ – rather than the training examples. RMM maximizes the margin while bounding the projections, in this case, the projections on the set $\{\mathbf{u}_1, \mathbf{u}_2, \ldots \mathbf{u}_{n_u}\}$ are bounded. Thus, the class of functions considered by the modified RMM is as follows:

Definition 2

$$\mathcal{H}_{E,D} := \{\mathbf{x} \rightarrow \mathbf{w}^\top \mathbf{x} | \frac{1}{2}\mathbf{w}^\top \mathbf{w} + \frac{D}{2}(\mathbf{w}^\top \mathbf{u}_i)^2 \leq E \ \forall 1 \leq i \leq n_u\},$$

where $D > 0$ trades off between large margin and small bound on the projections.

The above regularization scheme naturally suggests a third related function class which merely trades off between maximum absolute margin and a constraint on the average spread of the projections of all examples rather than a bound on individual projections:

Definition 3

$$\mathcal{G}_{E,D} := \{\mathbf{x} \rightarrow \mathbf{w}^\top \mathbf{x} | \frac{1}{2}\mathbf{w}^\top \mathbf{w} + \frac{D}{2n_u}\sum_{i=1}^{n_u}(\mathbf{w}^\top \mathbf{u}_i)^2 \leq E\}.$$

Note that this is closely related to the class of functions considered by $\mathbf{\Sigma}$-SVM.

The following lemmas elucidate some simple set theoretic relationships between the function classes and show that they form nested hypothesis spaces as is often encountered in structural risk minimization (SRM) [5]. The next section provides more in-depth Rademacher complexity estimates that relate the function classes.

Lemma 4 $\mathcal{G}_{E,0} = \mathcal{H}_{E,0} = \mathcal{F}_E$.

Lemma 5 $\mathcal{H}_{E,D} \subseteq \mathcal{G}_{E,D} \subseteq \mathcal{F}_E$.

Proof: Suppose $g(\cdot) \in \mathcal{H}_{E,D}$, then $\frac{1}{2}\mathbf{w}^\top\mathbf{w} + \frac{D}{2}(\mathbf{w}^\top\mathbf{u}_i)^2 \leq E \; \forall 1 \leq i \leq n_u$. Taking an average of the n_u constraints gives $\frac{1}{2}\mathbf{w}^\top\mathbf{w} + \frac{D}{2n_u}\sum_{i=1}^{n}(\mathbf{w}^\top\mathbf{u}_i)^2 \leq E$, thus $g(\cdot) \in \mathcal{G}_{E,D}$. This proves that $\mathcal{H}_{E,D} \subseteq \mathcal{G}_{E,D}$. Since $(\mathbf{w}^\top\mathbf{x}_i)^2 \geq 0$, we have $\mathcal{G}_{E,D} \subseteq \mathcal{F}_E$. ■

3 Rademacher complexity

This section will address the Rademacher complexity of the function classes in the previous section, in particular the empirical Rademacher complexity. Rademacher complexity measures richness of a class of real-valued functions with respect to a probability distribution [4, 2].

3.1 Preliminaries

Definition 6 *For a sample $S = \{\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_n\}$ generated by a distribution on $\mathbf{x}$ and a real valued function class $\mathcal{F}$ with domain $\mathbf{x}$, the empirical Rademacher complexity[1] of $\mathcal{F}$ is defined by:*

$$\hat{R}(\mathcal{F}) := \mathbf{E}_\sigma\left[\sup_{f\in\mathcal{F}}\left|\frac{2}{n}\sum_{i=1}^{n}\sigma_i f(\mathbf{x}_i)\right| \middle| S\right]$$

where $\sigma = \{\sigma_1, \ldots \sigma_n\}$ are independent random variables that take the values $+1$ or -1 with equal probability. Moreover, the Rademacher complexity of $\mathcal{F}$ is:

$$R(\mathcal{F}) := \mathbf{E}_S\left[\hat{R}(\mathcal{F})\right].$$

Essentially, Rademacher complexity quantifies how well a given function class can fit random labels. Keeping this quantity low reduces our ability to fit random labels and provides regularization for the learning problem.

Lemma 7 *Let $\mathcal{F}$ and $\mathcal{G}$ be classes of real functions. If $\mathcal{F} \subseteq \mathcal{G}$ then $\hat{R}(\mathcal{F}) \leq \hat{R}(\mathcal{G})$.*

The proof is straightforward, the supremum inside the definition of the empirical Rademacher complexity (Definition 6) is smaller when restricted to a smaller subset.

Corollary 8 $\hat{R}(\mathcal{H}_{E,D}) \leq \hat{R}(\mathcal{G}_{E,D}) \leq \hat{R}(\mathcal{F}_E)$.

3.2 Empirical Rademacher complexity of the function classes

In the rest of this section, we derive upper bounds on the empirical Rademacher complexities for the different function classes. These bounds provide insights on the regularization properties of the function classes. All the bounds are derived for the sample $S = \{\mathbf{x}_1, \mathbf{x}_2, \ldots \mathbf{x}_n\}$.

[1] We suppress the dependence of the empirical Rademacher complexity on n and S by writing $\hat{R}(\mathcal{F})$ for brevity.

Theorem 9 *The empirical Rademacher complexity on the sample S for the class $\mathcal{F}_E$ satisfies:*

$$\hat{R}(\mathcal{F}_E) \leq U_{\mathcal{F}_E} := \frac{2\sqrt{2E}}{n}\sqrt{\sum_{i=1}^{n}\mathbf{x}_i^\top\mathbf{x}_i}.$$

Proof:

$$\begin{aligned}
\hat{R}(\mathcal{F}_E) &= \mathbf{E}_\sigma\left[\sup_{f\in\mathcal{F}_E}\left|\frac{2}{n}\sum_{i=1}^{n}\sigma_i f(\mathbf{x}_i)\right|\right] \\
&= \mathbf{E}_\sigma\left[\max_{\frac{1}{2}\mathbf{w}^\top\mathbf{w}\leq E}\left|\mathbf{w}^\top\frac{2}{n}\sum_{i=1}^{n}\sigma_i\mathbf{x}_i\right|\right] \\
&= \frac{2}{n}\mathbf{E}_\sigma\left[\max_{||\mathbf{w}||\leq\sqrt{2E}}\left|\mathbf{w}^\top\sum_{i=1}^{n}\sigma_i\mathbf{x}_i\right|\right] \\
&\leq \frac{2\sqrt{2E}}{n}\mathbf{E}_\sigma\left[\left|\left|\sum_{i=1}^{n}\sigma_i\mathbf{x}_i\right|\right|\right] \\
&\leq \frac{2\sqrt{2E}}{n}\mathbf{E}_\sigma\left[\left(\sum_{i=1}^{n}\sigma_i\mathbf{x}_i^\top\sum_{j=1}^{n}\sigma_j\mathbf{x}_j\right)^{\frac{1}{2}}\right] \\
&\leq \frac{2\sqrt{2E}}{n}\left(\mathbf{E}_\sigma\left[\sum_{i,j=1}^{n}\sigma_i\sigma_j\mathbf{x}_i^\top\mathbf{x}_j\right]\right)^{\frac{1}{2}} \\
&= \frac{2\sqrt{2E}}{n}\sqrt{\sum_{i=1}^{n}\mathbf{x}_i^\top\mathbf{x}_i}.
\end{aligned}$$

In line three, the Cauchy-Schwarz inequality was applied. Subsequently, Jensen's inequality on the concave function $\sqrt{\cdot}$ is applied in line five. Finally, since σ_i and σ_j are random variables taking values $+1$ or -1 with equal probability, when $i \neq j$ $\mathbf{E}_\sigma[\sigma_i\sigma_j\mathbf{x}_i^\top\mathbf{x}_j] = 0$ and $\mathbf{E}_\sigma[\sigma_i\sigma_i\mathbf{x}_i^\top\mathbf{x}_i] = \mathbf{E}_\sigma[\mathbf{x}_i^\top\mathbf{x}_i] = \mathbf{x}_i^\top\mathbf{x}_i$. The result follows from the linearity of expectation. ■

Roughly speaking, by keeping E small, the ability to fit arbitrary labels is reduced. This is one way to motivate a maximum margin strategy. Note that $\sqrt{\sum_{i=1}^{n}\mathbf{x}_i^\top\mathbf{x}_i}$ is a measure of the spread of the data. However, most SVM formulations do not directly optimize this term. This motivates us to next consider the two new function classes that were defined as extensions to SVMs.

Theorem 10 *The empirical Rademacher complexity of the class $\mathcal{H}_{E,D}$, on the sample S, satisfies:*

$$\hat{R}(\mathcal{H}_{E,D}) \leq U_{\mathcal{H}_{E,D}} := \min_{\lambda\geq 0}\frac{1}{n}\sum_{i=1}^{n}\mathbf{x}_i^\top\mathbf{\Sigma}_{\lambda,D}^{-1}\mathbf{x}_i + \frac{2}{n}E\sum_{i=1}^{n_u}\lambda_i$$

where

$$\mathbf{\Sigma}_{\lambda,D} = \sum_{i=1}^{n_u}\lambda_i\mathbf{I} + D\sum_{i=1}^{n_u}\lambda_i\mathbf{u}_i\mathbf{u}_i^\top.$$

Proof:

$$\hat{R}(\mathcal{H}_{E,D}) = \mathbf{E}_\sigma \left[\sup_{f \in \mathcal{H}_{E,D}} \left| \frac{2}{n} \sum_{i=1}^{n} \sigma_i f(\mathbf{x}_i) \right| \right] \quad (1)$$
$$= \mathbf{E}_\sigma \left[\sup_{\mathbf{w}: \frac{1}{2}(\mathbf{w}^\top \mathbf{w} + D(\mathbf{w}^\top \mathbf{u}_i)^2) \leq E} \left| \frac{2}{n} \sum_{i=1}^{n} \sigma_i (\mathbf{w}^\top \mathbf{x}_i) \right| \right].$$

Consider the supremum inside the expectation. Depending on the sign of the term inside $|\cdot|$, it corresponds to either a maximization or a minimization. Without loss of generality, we consider the case of maximization. When a minimization is involved, the value of the objective still remains the same. The supremum is recovered by solving the following optimization problem:

$$\max_{\mathbf{w}} \ \mathbf{w}^\top \sum_{i=1}^{n} \sigma_i \mathbf{x}_i \quad (2)$$
$$\text{s.t. } \frac{1}{2}(\mathbf{w}^\top \mathbf{w} + D(\mathbf{w}^\top \mathbf{u}_i)^2) \leq E \quad \forall 1 \leq i \leq n_u.$$

The Lagrangian of the above optimization problem can be written as the following saddle problem involving convex minimization over the primal variables $\mathbf{w}$ with maximization over the non-negative Lagrange multipliers $\lambda_1, \lambda_2, \ldots \lambda_n$:

$$\mathcal{L}(\mathbf{w}, \lambda) = -\mathbf{w}^\top \sum_{i=1}^{n} \sigma_i \mathbf{x}_i + \sum_{i=1}^{n_u} \lambda_i \left(\frac{1}{2} \left(\mathbf{w}^\top \mathbf{w} + D(\mathbf{w}^\top \mathbf{u}_i)^2 \right) - E \right). \quad (3)$$

Differentiating (3) with respect to the primal variables gives:

$$\frac{\partial \mathcal{L}(\mathbf{w}, \lambda)}{\partial \mathbf{w}} = -\sum_{i=1}^{n} \sigma_i \mathbf{x}_i + \sum_{i=1}^{n_u} \lambda_i (\mathbf{w} + D \mathbf{u}_i \mathbf{u}_i^\top \mathbf{w}).$$

The optimum $\mathbf{w}$ is obtained by equating the above derivative to zero to give:

$$\left(\sum_{i=1}^{n_u} \lambda_i \mathbf{I} + D \sum_{i=1}^{n_u} \lambda_i \mathbf{u}_i \mathbf{u}_i^\top \right) \mathbf{w} = \sum_{i=1}^{n} \sigma_i \mathbf{x}_i.$$

Using the definition $\mathbf{\Sigma}_{\lambda,D} := \sum_{i=1}^{n_u} \lambda_i \mathbf{I} + D \sum_{i=1}^{n_u} \lambda_i \mathbf{u}_i \mathbf{u}_i^\top$; we now have:

$$\mathbf{w} = \mathbf{\Sigma}_{\lambda,D}^{-1} \sum_{i=1}^{n} \sigma_i \mathbf{x}_i. \quad (4)$$

Substituting the expression for $\mathbf{w}$ in (3):

$$\begin{aligned}
&-\mathbf{w}^\top \sum_{i=1}^{n} \sigma_i \mathbf{x}_i + \sum_{i=1}^{n_u} \lambda_i \left(\frac{1}{2} \left(\mathbf{w}^\top \mathbf{w} + D(\mathbf{w}^\top \mathbf{u}_i)^2 \right) - E \right) \\
&= -\sum_{i=1}^{n} \sigma_i \mathbf{x}_i^\top \mathbf{\Sigma}_{\lambda,D}^{-1} \sum_{i=1}^{n} \sigma_i \mathbf{x}_i - E \sum_{i=1}^{n_u} \lambda_i \\
&\quad + \frac{1}{2} \sum_{i=1}^{n_u} \lambda_i \mathbf{w}^\top (\mathbf{I} + D \mathbf{u}_i \mathbf{u}_i^\top) \mathbf{w} \\
&= -\sum_{i=1}^{n} \sigma_i \mathbf{x}_i^\top \mathbf{\Sigma}_{\lambda,D}^{-1} \sum_{i=1}^{n} \sigma_i \mathbf{x}_i - E \sum_{i=1}^{n_u} \lambda_i \\
&\quad + \frac{1}{2} \mathbf{w}^\top \left(\sum_{i=1}^{n_u} \lambda_i \mathbf{I} + D \sum_{i=1}^{n_u} \lambda_i \mathbf{u}_i \mathbf{u}_i^\top \right) \mathbf{w} \\
&= -\sum_{i=1}^{n} \sigma_i \mathbf{x}_i^\top \mathbf{\Sigma}_{\lambda,D}^{-1} \sum_{i=1}^{n} \sigma_i \mathbf{x}_i - E \sum_{i=1}^{n_u} \lambda_i \\
&\quad + \frac{1}{2} \sum_{i=1}^{n} \sigma_i \mathbf{x}_i^\top \mathbf{\Sigma}_{\lambda,D}^{-1} \sum_{i=1}^{n} \sigma_i \mathbf{x}_i \\
&= -\frac{1}{2} \sum_{i=1}^{n} \sigma_i \mathbf{x}_i^\top \mathbf{\Sigma}_{\lambda,D}^{-1} \sum_{i=1}^{n} \sigma_i \mathbf{x}_i - E \sum_{i=1}^{n_u} \lambda_i.
\end{aligned}$$

Thus the dual of the formulation (2) is given by:

$$\min_{\lambda \geq 0} \ \frac{1}{2} \sum_{i=1}^{n} \sigma_i \mathbf{x}_i^\top \mathbf{\Sigma}_{\lambda,D}^{-1} \sum_{i=1}^{n} \sigma_i \mathbf{x}_i + E \sum_{i=1}^{n_u} \lambda_i. \quad (5)$$

Now, we derive an upper bound on the empirical Rademacher complexity $\hat{R}(\mathcal{H}_{E,D})$:

$$\begin{aligned}
&\hat{R}(\mathcal{H}_{E,D}) \\
&= \mathbf{E}_\sigma \left[\sup_{\mathbf{w}: \frac{1}{2}(\mathbf{w}^\top \mathbf{w} + D(\mathbf{w}^\top \mathbf{u}_i)^2) \leq E} \left| \frac{2}{n} \sum_{i=1}^{n} \sigma_i (\mathbf{w}^\top \mathbf{x}_i) \right| \right] \\
&= \frac{2}{n} \mathbf{E}_\sigma \left[\min_{\lambda \geq 0} \ \frac{1}{2} \sum_{i=1}^{n} \sigma_i \mathbf{x}_i^\top \mathbf{\Sigma}_{\lambda,D}^{-1} \sum_{j=1}^{n} \sigma_j \mathbf{x}_j + E \sum_{i=1}^{n_u} \lambda_i \right] \\
&\leq \min_{\lambda \geq 0} \frac{2}{n} \mathbf{E}_\sigma \left[\frac{1}{2} \sum_{i=1}^{n} \sigma_i \mathbf{x}_i^\top \mathbf{\Sigma}_{\lambda,D}^{-1} \sum_{j=1}^{n} \sigma_j \mathbf{x}_j + E \sum_{i=1}^{n_u} \lambda_i \right] \\
&\leq \min_{\lambda \geq 0} \frac{2}{n} \mathbf{E}_\sigma \left[\frac{1}{2} \sum_{i,j=1} \sigma_i \sigma_j \mathbf{x}_i^\top \mathbf{\Sigma}_{\lambda,D}^{-1} \mathbf{x}_j + E \sum_{i=1}^{n_u} \lambda_i \right] \\
&\leq \min_{\lambda \geq 0} \frac{1}{n} \sum_{i=1}^{n} \mathbf{x}_i^\top \mathbf{\Sigma}_{\lambda,D}^{-1} \mathbf{x}_i + \frac{2}{n} E \sum_{i=1}^{n_u} \lambda_i. \quad (6)
\end{aligned}$$

To go from the first line to the second, we use the fact that the primal (2) and the dual (5) objectives are equal at the optimum. On line two, the expectation is over the minimizers over λ; clearly, this is less than first taking the expectation and then minimizing over λ in line three. On line four, simply recycle the arguments used in Theorem 9 to handle the expectation over σ. ■

Note that the upper bound $U_{\mathcal{H}_{E,D}}$ is not a closed form expression in the general case but is possible to evaluate in polynomial time using semi-definite programming. Using Schur's complement lemma (Appendix A) the minimization (6) can be expressed as the following semi-definite optimization [3]:

$$\min_{\lambda \geq 0, t} \frac{1}{n}\sum_{i=1}^{n} t_i + \frac{2E}{n}\sum_{i=1}^{n_u}\lambda_i \qquad (7)$$
$$\text{s.t.} \begin{bmatrix} \sum_{j=1}^{n_u}\lambda_j(\mathbf{I} + D\mathbf{u}_j\mathbf{u}_j^\top) & \mathbf{x}_i \\ \mathbf{x}_i^\top & t_i \end{bmatrix} \succeq 0 \quad \forall 1 \leq i \leq n.$$

It is interesting to note that even though it involved different derivation steps, the upper bound in Theorem 10 (namely, $U_{\mathcal{H}_{E,D}}$) is no looser than the upper bound in Theorem 9 (namely, $U_{\mathcal{F}_E}$) when $D = 0$; they exactly coincide in that case. The following theorem makes this clear:

Theorem 11 *When $D = 0$, $U_{\mathcal{H}_{E,D}} = U_{\mathcal{F}_E}$.*

Proof:

$$U_{\mathcal{H}_{E,0}} = \min_{\lambda \geq 0} \frac{1}{n}\sum_{i=1}^{n}\mathbf{x}_i^\top \boldsymbol{\Sigma}_{\lambda,0}^{-1}\mathbf{x}_i + \frac{2}{n}E\sum_{i=1}^{n_u}\lambda_i$$
$$= \min_{\lambda \geq 0} \frac{1}{n}\frac{\sum_{i=1}^{n}\mathbf{x}_i^\top\mathbf{x}_i}{\sum_{i=1}^{n_u}\lambda_i} + \frac{2}{n}E\sum_{i=1}^{n_u}\lambda_i.$$

In non-trivial cases, defining $\tau := \sum_{i=1}^{n_u}\lambda_i$, we have:

$$U_{\mathcal{H}_{E,0}} = \min_{\tau \geq 0} \frac{1}{n\tau}\sum_{i=1}^{n}\mathbf{x}_i^\top\mathbf{x}_i + \frac{2}{n}E\tau.$$

Differentiating the above expression with respect to τ and selecting the positive solution, we get:

$$\tau = \frac{\sqrt{\sum_{i=1}^{n}\mathbf{x}_i^\top\mathbf{x}_i}}{\sqrt{2E}}.$$

This τ when substituted back in the expression for $U_{\mathcal{H}_{E,0}}$ gives $U_{\mathcal{F}_E}$. ■

As the value of D is increased, our bound on the empirical Rademacher complexity decreases. However, this decrease is not merely due to a change in the scale of the extra term in $\boldsymbol{\Sigma}_{\lambda,D}$ (which would be of limited value in practice). Instead, increasing D provides increased flexibility in optimizing over the Lagrange multipliers λ which affect the shape of the matrix $\sum_{i=1}^{n}\lambda_i\mathbf{x}_i\mathbf{x}_i^\top$ in addition to scaling it. Thus, the empirical Rademacher complexity is not necessarily simply linear in D.

At the optimum, (4) suggests that the classifier must be

$$\mathbf{w}^* = \Sigma_{\lambda^*,D}^{-1}\sum_{i=1}^{n}\sigma_i\mathbf{x}_i.$$

From (1), we have:

$$\hat{R}(\mathcal{H}_{E,D}) = \frac{2}{n}\mathbf{E}_\sigma\left[\sum_{i=1}^{n}\sigma_i\mathbf{x}_i\boldsymbol{\Sigma}_{\lambda^*,D}^{-1}\sum_{i=1}^{n}\sigma_i\mathbf{x}_i\right].$$

Note that $\boldsymbol{\Sigma}_{\lambda,D}^*$ is the optimal solution obtained from (5). Formulation (5) is precisely trying to minimize the term inside the expectation above by changing the shape of $\boldsymbol{\Sigma}_{\lambda,D}$ using λ. Not only is the magnitude of D relevant but the overall shape of $\boldsymbol{\Sigma}_{\lambda,D}$ also plays a key role in reducing the empirical Rademacher complexity.

Theorem 12 *The empirical Rademacher complexity of the class $\mathcal{G}_{E,D}$ on the sample S satisfies:*

$$\hat{R}(\mathcal{G}_{E,D}) \leq U_{\mathcal{G}_{E,D}} := \frac{2\sqrt{2E}}{n}\left(\sum_{i=1}^{n}\mathbf{x}_i^\top\boldsymbol{\Sigma}_D^{-1}\mathbf{x}_i\right)^{\frac{1}{2}}$$

where

$$\boldsymbol{\Sigma}_D = \mathbf{I} + \frac{D}{n_u}\sum_{i=1}^{n_u}\mathbf{u}_i\mathbf{u}_i^\top.$$

Proof: Following steps similar to those in Theorem 12, we start with the optimization:

$$\max_{\mathbf{w}} \ \mathbf{w}^\top\sum_{i=1}^{n}\sigma_i\mathbf{x}_i \qquad (8)$$
$$\text{s.t.} \ \frac{1}{2n_u}\sum_{i=1}^{n_u}\left(\mathbf{w}^\top\mathbf{w} + D(\mathbf{w}^\top\mathbf{u}_i)^2\right) \leq E.$$

We write the Lagrangian of the above problem as:

$$\mathcal{L}(\mathbf{w},\lambda) = -\mathbf{w}^\top\sum_{i=1}^{n}\sigma_i\mathbf{x}_i + \lambda\left(\frac{1}{2}\mathbf{w}^\top\boldsymbol{\Sigma}_D\mathbf{w} - E\right),$$

where λ is a non-negative Lagrange multiplier. Differentiating and equating the partial derivative with respect to $\mathbf{w}$ to zero in the Lagrangian, we get:

$$\mathbf{w} = \frac{1}{\lambda}\boldsymbol{\Sigma}_D^{-1}\sum_{i=1}^{n}\sigma_i\mathbf{x}_i. \qquad (9)$$

Since we are optimizing a linear objective over a single quadratic constraint, at the optimum, the quadratic constraint becomes tight. Using this fact we can write:

$$\frac{1}{2}\mathbf{w}^\top\boldsymbol{\Sigma}_D\mathbf{w} = E.$$

Substituting $\mathbf{w}$ from (9) in the above produces:

$$\lambda = \frac{1}{\sqrt{2E}}\sqrt{\sum_{i=1}^{n}\sigma_i\mathbf{x}_i^\top\boldsymbol{\Sigma}_D^{-1}\sum_{i=1}^{n}\sigma_i\mathbf{x}_i}.$$

Substituting this λ back in (9) and using that $\mathbf{w}$ in the objective (8) yields:

$$\sqrt{2E}\sqrt{\sum_{i=1}^{n}\sigma_i\mathbf{x}_i^\top\boldsymbol{\Sigma}_D^{-1}\sum_{i=1}^{n}\sigma_i\mathbf{x}_i}.$$

Thus, the empirical Rademacher complexity is given by:

$$\begin{aligned}
&\frac{2}{n}\mathbf{E}_\sigma\left[\sqrt{2E}\sqrt{\sum_{i=1}^{n}\sigma_i\mathbf{x}_i^\top\mathbf{\Sigma}_D^{-1}\sum_{i=1}^{n}\sigma_i\mathbf{x}_i}\right]\\
\leq&\frac{2\sqrt{2E}}{n}\left(\mathbf{E}_\sigma\left[\sum_{i=1}^{n}\sigma_i\mathbf{x}_i^\top\mathbf{\Sigma}_D^{-1}\sum_{i=1}^{n}\sigma_i\mathbf{x}_i\right]\right)^{\frac{1}{2}}\\
=&\frac{2\sqrt{2E}}{n}\left(\sum_{i=1}^{n}\mathbf{x}_i^\top\mathbf{\Sigma}_D^{-1}\mathbf{x}_i\right)^{\frac{1}{2}}.
\end{aligned}$$

The second line follows from Jensen's inequality on the concave function $\sqrt{\cdot}$. ■

As we saw before, $U_{\mathcal{H}_{0,E}}$ coincides with $U_{\mathcal{F}_E}$. In addition, some of the properties of $U_{\mathcal{G}_{D,E}}$ also carry over to $U_{\mathcal{H}_{D,E}}$. Note, however, that the term $\sum_{i=1}^{n_u}\lambda_i\mathbf{u}_i\mathbf{u}_i^\top$ appearing in $\mathbf{\Sigma}_{D,\lambda}$ is more flexible than the term $\sum_{i=1}^{n}\mathbf{u}_i\mathbf{u}_i^\top$ in $\mathbf{\Sigma}_D$ since the matrix $\mathbf{\Sigma}_D$ only undergoes scaling as D is varied.

3.3 Relation to the actual Rademacher complexity

Note that, by definition 6, the empirical Rademacher complexity of a function class is dependent on the data (sample S). In many cases, it is not possible to give exact expressions for the Rademacher complexity since the underlying distribution over the data is unknown. However, it is possible to give probabilistic upper bounds on the Rademacher complexity. Since the Rademacher complexity is the expectation of its empirical estimate over the data, by a straightforward application of McDiarmid's inequality, it is possible to show the following:

Lemma 13 *Fix $\delta \in (0,1)$, with probability at least $1-\delta$ over draws of the samples S the following holds for any function class $\mathcal{F}$:*

$$R(\mathcal{F}) \leq \hat{R}(\mathcal{F}) + 2\sqrt{\frac{\ln(2/\delta)}{2n}}.$$

In other words, the probability of one-sided deviation[2] between $R(\mathcal{F})$ and $\hat{R}(\mathcal{F})$ drops off exponentially with n. This suggests that even though we have used empirical estimates of the Rademacher complexity, they are not too far from their actual values with high probability.

4 Generalization bounds

This section presents generalization bounds for the three different function classes. A generic bound for the general case is first derived and then applied to the three cases. The derivation largely follows the approach of [4] and therefore details will be omitted in this article.

[2]It is possible to state a similar result for the two-sided absolute deviation but we have used a one sided result since we are primarily interested in upper bounds.

Recall the theorem from [4] that leverages the empirical Rademacher complexity to provide a generalization bound.

Theorem 14 *Let $\mathcal{F}$ be a class of functions mapping Z to $[0,1]$; let $\{\mathbf{z}_1,\ldots,\mathbf{z}_n\}$ be drawn from the domain Z independently and identically distributed (iid) according to a probability distribution $\mathcal{D}$. Then for any fixed $\delta \in (0,1)$, the following bound holds for any $f \in \mathcal{F}$ with probability at least $1-\delta$ over random draws of a set of samples of size n:*

$$\boldsymbol{E}_{\mathcal{D}}[f(\mathbf{z})] \leq \hat{\boldsymbol{E}}[f(\mathbf{z})] + \hat{R}(\mathcal{F}) + 3\sqrt{\frac{\ln(2/\delta)}{2n}}.$$

Definition 15 *The Heaviside function, $Q:\mathbb{R}\to\{0,1\}$ is defined as:*

$$Q(s) := \begin{cases} 1 & \text{if } s > 0,\\ 0 & \text{otherwise.}\end{cases}$$

For a function $g:\mathbb{R}^m\to\mathbb{R}$, $Q(-yg(\mathbf{x}))$ is an indicator of whether the function $g(\cdot)$ predicts the right label for the example $(\mathbf{x},y)\in\mathbb{R}^m\times\{\pm1\}$. So, if we let $f(\mathbf{x},y) = -yg(\mathbf{x})$, we have:

$$\boldsymbol{E}_{\mathcal{D}}[Q(f(\mathbf{x},y))] = \boldsymbol{E}_{\mathcal{D}}[Q(-yg(\mathbf{x}))] = \Pr_{\mathcal{D}}[y \neq \text{sign}(g(\mathbf{x}))],$$

which is exactly the generalization error that we would like to bound.

Theorem 16 *Fix $\gamma > 0$, let $\mathcal{F}$ be the class of functions from $\mathbb{R}^m\times\{\pm1\}\to\mathbb{R}$ given by $f(\mathbf{x},y) = -yg(\mathbf{x})$. Let $\{(\mathbf{x}_1,y_1),\ldots,(\mathbf{x}_n,y_n)\}$ be drawn iid from a probability distribution $\mathcal{D}$. Then, with probability at least $1-\delta$ over the samples of size n, the following bound holds:*

$$\begin{aligned}
&\Pr_{\mathcal{D}}[y \neq sign(g(\mathbf{x}))]\\
&\leq \frac{1}{n}\sum_{i=1}^{n}\xi_i + \frac{2}{\gamma}\hat{R}(\mathcal{F}) + 3\sqrt{\frac{\ln(2/\delta)}{2n}},
\end{aligned} \tag{10}$$

where $\xi_i = \max(0, 1-y_ig(\mathbf{x}_i))$ are the so-called slack variables.

Proof: We first define the function $\mathcal{A}:\mathbb{R}\to[0,1]$ as below:

$$\mathcal{A}(s) := \begin{cases} 1, & \text{if } s > 0,\\ 1+\frac{s}{\gamma} & \text{if } -\gamma \leq s \leq 0,\\ 0 & \text{otherwise.}\end{cases}$$

First, note that $Q(f(\mathbf{x},y)) \leq \mathcal{A}(f(\mathbf{x},y))$. Thus,

$$\boldsymbol{E}_{\mathcal{D}}[Q(f(\mathbf{x},y)) - 1] \leq \boldsymbol{E}_{\mathcal{D}}[\mathcal{A}(f(\mathbf{x},y)) - 1].$$

Theorem 14 can now be applied to $\boldsymbol{E}_{\mathcal{D}}[\mathcal{A}(f(\mathbf{x},y)) - 1]$ to get:

$$\begin{aligned}
&\boldsymbol{E}_{\mathcal{D}}[Q(f(\mathbf{x},y)) - 1] \leq \boldsymbol{E}_{\mathcal{D}}[\mathcal{A}(f(\mathbf{x},y)) - 1]\\
&\leq \hat{\boldsymbol{E}}[\mathcal{A}(f(\mathbf{x},y)) - 1] + \hat{R}((\mathcal{A}-1)\circ\mathcal{F}) + 3\sqrt{\frac{\ln(2/\delta)}{2n}}.
\end{aligned}$$

The above simplifies as:

$$\boldsymbol{E}_{\mathcal{D}}[Q(f(\mathbf{x},y))]$$
$$\leq \hat{\boldsymbol{E}}[\mathcal{A}(f(\mathbf{x},y))] \quad + \hat{R}((\mathcal{A}-1)\circ\mathcal{F}) + 3\sqrt{\frac{\ln(2/\delta)}{2n}}.$$

However, it is easy to see that:

$$\hat{\boldsymbol{E}}[\mathcal{A}(f(\mathbf{x},y))] = \sum_{i=1}^{n}\mathcal{A}(f(\mathbf{x}_i,y_i)) \leq \frac{1}{n}\sum_{i=1}^{n}\xi_i.$$

Thus the only quantity that remains to be bounded is $\hat{R}((\mathcal{A}-1)\circ\mathcal{F})$. It can be shown that this quantity satisfies the theorem in Appendix B with $L = \frac{1}{\gamma}$, to give:

$$\hat{R}((\mathcal{A}-1)\circ\mathcal{F}) \leq \frac{2}{\gamma}\hat{R}(\mathcal{F}).$$

■

We can now substitute the upper bounds that we derived in Section 3 , namely: $U_{\mathcal{F}_E}, U_{\mathcal{G}_{E,D}}$ and $U_{\mathcal{H}_{E,D}}$, in the above theorem to get the corresponding generalization bounds for the function classes of interest. In contrast to the bound for the function class $\mathcal{F}_E$, the other two bounds provide a more explicit role for the spread of the data.

There is a trade-off between the average of the slack variables and the empirical Rademacher term (the complexity term) in the generalization bound. In many problems (for example in high dimensional feature spaces or when nonlinear kernels are used) the average slack variables can be small if training data can be easily separated. In such situations, reducing the complexity term can significantly drive down the bound.

References

[1] A. Ambroladze and J. Shawe-Taylor. Complexity of pattern classes and Lipschitz property. In *Algorithmic Learning Theory*, pages 181–193, 2004.

[2] O. Bousquet, S. Boucheron, and G. Lugosi. *Introduction to Statistical Learning Theory*, volume Lecture Notes in Artificial Intelligence 3176, pages 169–207. Springer, Heidelberg, Germany, 2004.

[3] S. Boyd and L. Vandenberghe. *Convex Optimization*. Cambridge University Press, 2003.

[4] J. Shawe-Taylor and N. Cristianini. *Kernel Methods for Pattern Analysis*. Cambridge University Press, 2004.

[5] V. Vapnik. *The Nature of Statistical Learning Theory*. Springer Verlag, New York, 1995.

A Schur's complement lemma

Let $\mathbf{A} \in \mathbb{R}^{m\times m}, \mathbf{B} \in \mathbb{R}^{m\times 1}, C \in \mathbb{R}$; further let $\mathbf{A} \succ 0$, then:

$$\begin{bmatrix} \mathbf{A} & \mathbf{B} \\ \mathbf{B}^\top & C \end{bmatrix} \succeq 0 \text{ if and only if } C - \mathbf{B}^\top\mathbf{A}^{-1}\mathbf{B} \geq 0.$$

B A property of the empirical Rademacher complexity

A function $f : \mathbb{R} \to \mathbb{R}$ is Lipschitz with constant L if: $|f(x) - f(y)| \leq L|x-y|$, for any $x, y \in \mathbb{R}$.

Theorem 17 *If $\mathcal{A} : \mathbb{R} \to \mathbb{R}$ is Lipschitz with constant L and if $\mathcal{A}(0) = 0$ then $\hat{R}(\mathcal{A}\circ\mathcal{F}) \leq 2L\hat{R}(\mathcal{F})$.*

A proof can be found in [1].