14 Algorithmic Stability

In chapters 2–5 and several subsequent chapters, we presented a variety of generalization bounds based on different measures of the complexity of the hypothesis set $\mathcal{H}$ used for learning, including the Rademacher complexity, the growth function, and the VC-dimension. These bounds ignore the specific algorithm used, that is, they hold for any algorithm using $\mathcal{H}$ as a hypothesis set.

One may ask if an analysis of the properties of a specific algorithm could lead to finer guarantees. Such an algorithm-dependent analysis could have the benefit of a more informative guarantee. On the other hand, it could be inapplicable to other algorithms using the same hypothesis set. Alternatively, as we shall see in this chapter, a more general property of the learning algorithm could be used to incorporate algorithm-specific properties while extending the applicability of the analysis to other learning algorithms with similar properties.

This chapter uses the property of *algorithmic stability* to derive *algorithm-dependent* learning guarantees. We first present a generalization bound for any algorithm that is sufficiently stable. Then, we show that the wide class of kernel-based regularization algorithms enjoys this property and derive a general upper bound on their stability coefficient. Finally, we illustrate the application of these results to the analysis of several algorithms both in the regression and classification settings, including kernel ridge regression (KRR), SVR, and SVMs.

14.1 Definitions

We start by introducing the notation and definitions relevant to our analysis of algorithmic stability. We denote by z a labeled example $(x, y) \in \mathcal{X} \times \mathcal{Y}$. The hypotheses h we consider map $\mathcal{X}$ to a set $\mathcal{Y}'$ sometimes different from $\mathcal{Y}$. In particular, for classification, we may have $\mathcal{Y} = \{-1, +1\}$ while the hypothesis h learned takes values in $\mathbb{R}$. The loss functions L we consider are therefore defined over $\mathcal{Y}' \times \mathcal{Y}$, with $\mathcal{Y}' = \mathcal{Y}$ in most cases. For a loss function $L \colon \mathcal{Y}' \times \mathcal{Y} \to \mathbb{R}_+$, we denote the loss of a hypothesis h at point z by $L_z(h) = L(h(x), y)$. We denote by $\mathcal{D}$ the distribution according to which samples are drawn and by $\mathcal{H}$ the hypothesis set. The empirical error or loss of $h \in \mathcal{H}$ on a sample $S = (z_1, \ldots, z_m)$ and its generalization error are defined, respectively, by

$$\widehat{R}_S(h) = \frac{1}{m} \sum_{i=1}^{m} L_{z_i}(h) \quad \text{and} \quad R(h) = \mathop{\mathbb{E}}_{z \sim \mathcal{D}}[L_z(h)].$$

Given an algorithm $\mathcal{A}$, we denote by h_S the hypothesis $h_S \in \mathcal{H}$ returned by $\mathcal{A}$ when trained on sample S. We will say that the loss function L is bounded by $M \geq 0$ if for all $h \in \mathcal{H}$ and $z \in \mathcal{X} \times \mathcal{Y}$, $L_z(h) \leq M$. For the results presented in this chapter, a weaker condition suffices, namely that $L_z(h_S) \leq M$ for all hypotheses h_S returned by the algorithm $\mathcal{A}$.

We are now able to define the notion of uniform stability, the algorithmic property used in the analyses of this chapter.

Definition 14.1 (Uniform stability) *Let S and S' be any two training samples that differ by a single point. Then, a learning algorithm $\mathcal{A}$ is* uniformly β-stable *if the hypotheses it returns when trained on any such samples S and S' satisfy*

$$\forall z \in \mathcal{Z}, \quad |L_z(h_S) - L_z(h_{S'})| \leq \beta.$$

The smallest such β satisfying this inequality is called the stability coefficient *of $\mathcal{A}$.*

In other words, when $\mathcal{A}$ is trained on two similar training sets, the losses incurred by the corresponding hypotheses returned by $\mathcal{A}$ should not differ by more than β. Note that a uniformly β-stable algorithm is often referred to as being *β-stable* or even just *stable* (for some unspecified β). In general, the coefficient β depends on the sample size m. We will see in section 14.2 that $\beta = o(1/\sqrt{m})$ is necessary for the convergence of the stability-based learning bounds presented in this chapter. In section 14.3, we will show that a more favorable condition holds, that is, $\beta = O(1/m)$, for a wide family of algorithms.

14.2 Stability-based generalization guarantee

In this section, we show that exponential bounds can be derived for the generalization error of stable learning algorithms. The main result is presented in theorem 14.2.

Theorem 14.2 *Assume that the loss function L is bounded by $M \geq 0$. Let $\mathcal{A}$ be a β-stable learning algorithm and let S be a sample of m points drawn i.i.d. according* *to distribution* $\mathcal{D}$. *Then, with probability at least* $1-\delta$ *over the sample* S *drawn, the following holds:*

$$R(h_S) \le \widehat{R}_S(h_S) + \beta + (2m\beta + M)\sqrt{\frac{\log\frac{1}{\delta}}{2m}}.$$

Proof: The proof is based on the application of McDiarmid's inequality (theorem D.8) to the function Φ defined for all samples S by $\Phi(S) = R(h_S) - \widehat{R}_S(h_S)$. Let S' be another sample of size m with points drawn i.i.d. according to $\mathcal{D}$ that differs from S by exactly one point. We denote that point by z_m in S, z'_m in S', i.e.,

$$S = (z_1, \ldots, z_{m-1}, z_m) \quad \text{and} \quad S' = (z_1, \ldots, z_{m-1}, z'_m).$$

By definition of Φ, the following inequality holds:

$$|\Phi(S') - \Phi(S)| \le |R(h_{S'}) - R(h_S)| + |\widehat{R}_{S'}(h_{S'}) - \widehat{R}_S(h_S)|. \tag{14.1}$$

We bound each of these two terms separately. By the β-stability of $\mathcal{A}$, we have

$$|R(h_S) - R(h_{S'})| = |\mathop{\mathbb{E}}_z[L_z(h_S)] - \mathop{\mathbb{E}}_z[L_z(h_{S'})]| \le \mathop{\mathbb{E}}_z[|L_z(h_S) - L_z(h_{S'})|] \le \beta.$$

Using the boundedness of L along with β-stability of $\mathcal{A}$, we also have

$$\begin{aligned} |\widehat{R}_S(h_S) - \widehat{R}_{S'}(h_{S'})| &= \frac{1}{m}\left|\left(\sum_{i=1}^{m-1} L_{z_i}(h_S) - L_{z_i}(h_{S'})\right) + L_{z_m}(h_S) - L_{z'_m}(h_{S'})\right| \\ &\le \frac{1}{m}\left[\left(\sum_{i=1}^{m-1} |L_{z_i}(h_S) - L_{z_i}(h_{S'})|\right) + |L_{z_m}(h_S) - L_{z'_m}(h_{S'})|\right] \\ &\le \frac{m-1}{m}\beta + \frac{M}{m} \le \beta + \frac{M}{m}. \end{aligned}$$

Thus, in view of (14.1), Φ satisfies the condition $|\Phi(S) - \Phi(S')| \le 2\beta + \frac{M}{m}$. By applying McDiarmid's inequality to $\Phi(S)$, we can bound the deviation of Φ from its mean as

$$\mathbb{P}\left[\Phi(S) \ge \epsilon + \mathop{\mathbb{E}}_S[\Phi(S)]\right] \le \exp\left(\frac{-2m\epsilon^2}{(2m\beta + M)^2}\right),$$

or, equivalently, with probability $1-\delta$,

$$\Phi(S) < \epsilon + \mathop{\mathbb{E}}_S[\Phi(S)], \tag{14.2}$$

where $\delta = \exp\left(\frac{-2m\epsilon^2}{(2m\beta+M)^2}\right)$. If we solve for ϵ in this expression for δ, plug into (14.2) and rearrange terms, then, with probability $1-\delta$, we have

$$\Phi(S) \le \mathop{\mathbb{E}}_{S\sim\mathcal{D}^m}[\Phi(S)] + (2m\beta + M)\sqrt{\frac{\log\frac{1}{\delta}}{2m}}. \tag{14.3}$$

We now bound the expectation term, first noting that by linearity of expectation $\mathbb{E}_S[\Phi(S)] = \mathbb{E}_S[R(h_S)] - \mathbb{E}_S[\widehat{R}_S(h_S)]$. By definition of the generalization error,

$$\underset{S\sim\mathcal{D}^m}{\mathbb{E}}[R(h_S)] = \underset{S\sim\mathcal{D}^m}{\mathbb{E}}\Big[\underset{z\sim\mathcal{D}}{\mathbb{E}}[L_z(h_S)]\Big] = \underset{S,z\sim\mathcal{D}^{m+1}}{\mathbb{E}}[L_z(h_S)]. \tag{14.4}$$

By the linearity of expectation,

$$\underset{S\sim\mathcal{D}^m}{\mathbb{E}}[\widehat{R}_S(h_S)] = \frac{1}{m}\sum_{i=1}^{m}\underset{S\sim\mathcal{D}^m}{\mathbb{E}}[L_{z_i}(h_S)] = \underset{S\sim\mathcal{D}^m}{\mathbb{E}}[L_{z_1}(h_S)], \tag{14.5}$$

where the second equality follows from the fact that the z_i are drawn i.i.d. and thus the expectations $\mathbb{E}_{S\sim\mathcal{D}^m}[L_{z_i}(h_S)]$, $i \in [m]$, are all equal. The last expression in (14.5) is the expected loss of a hypothesis on one of its training points. We can rewrite it as $\mathbb{E}_{S\sim D^m}[L_{z_1}(h_S)] = \mathbb{E}_{S,z\sim\mathcal{D}^{m+1}}[L_z(h_{S'})]$, where S' is a sample of m points containing z extracted from the $m+1$ points formed by S and z. Thus, in view of (14.4) and by the β-stability of $\mathcal{A}$, it follows that

$$\begin{aligned}
\big|\underset{S\sim\mathcal{D}^m}{\mathbb{E}}[\Phi(S)]\big| &= \Big|\underset{S,z\sim\mathcal{D}^{m+1}}{\mathbb{E}}[L_z(h_S)] - \underset{S,z\sim\mathcal{D}^{m+1}}{\mathbb{E}}[L_z(h_{S'})]\Big| \\
&\leq \underset{S,z\sim\mathcal{D}^{m+1}}{\mathbb{E}}\big[|L_z(h_S) - L_z(h_{S'})|\big] \\
&\leq \underset{S,z\sim\mathcal{D}^{m+1}}{\mathbb{E}}[\beta] = \beta.
\end{aligned}$$

We can thus replace $\mathbb{E}_S[\Phi(S)]$ by β in (14.3), which completes the proof. □

The bound of the theorem converges for $(m\beta)/\sqrt{m} = o(1)$, that is $\beta = o(1/\sqrt{m})$. In particular, when the stability coefficient β is in $O(1/m)$, the theorem guarantees that $R(h_S) - \widehat{R}_S(h_S) = O(1/\sqrt{m})$ with high probability. In the next section, we show that kernel-based regularization algorithms precisely admit this property under some general assumptions.

14.3 Stability of kernel-based regularization algorithms

Let K be a positive definite symmetric kernel, $\mathbb{H}$ the reproducing kernel Hilbert space associated to K, and $\|\cdot\|_K$ the norm induced by K in $\mathbb{H}$. A kernel-based regularization algorithm is defined by the minimization over $\mathbb{H}$ of an objective function F_S based on a training sample $S = (z_1, \ldots, z_m)$ and defined for all $h \in \mathbb{H}$ by:

$$F_S(h) = \widehat{R}_S(h) + \lambda\|h\|_K^2. \tag{14.6}$$

In this equation, $\widehat{R}_S(h) = \frac{1}{m}\sum_{i=1}^{m} L_{z_i}(h)$ is the empirical error of hypothesis h with respect to a loss function L and $\lambda \geq 0$ a trade-off parameter balancing the emphasis on the empirical error versus the regularization term $\|h\|_K^2$. The hypothesis set $\mathcal{H}$

is the subset of $\mathbb{H}$ formed by the hypotheses possibly returned by the algorithm. Algorithms such as KRR, SVR and SVMs all fall under this general model.

We first introduce some definitions and tools needed for a general proof of an upper bound on the stability coefficient of kernel-based regularization algorithms. Our analysis will assume that the loss function L is convex and that it further verifies the following Lipschitz-like smoothness condition.

Definition 14.3 (σ-admissibility) *A loss function L is* σ-admissible *with respect to the hypothesis class $\mathcal{H}$ if there exists $\sigma \in \mathbb{R}_+$ such that for any two hypotheses $h, h' \in \mathcal{H}$ and for all $(x, y) \in \mathcal{X} \times \mathcal{Y}$,*

$$|L(h'(x), y) - L(h(x), y)| \leq \sigma |h'(x) - h(x)|. \tag{14.7}$$

This assumption holds for the quadratic loss and most other loss functions where the hypothesis set and the set of output labels are bounded by some $M \in \mathbb{R}_+$: $\forall h \in \mathcal{H}, \forall x \in \mathcal{X}, |h(x)| \leq M$ and $\forall y \in \mathcal{Y}, |y| \leq M$.

We will use the notion of *Bregman divergence*, B_F which can be defined for any convex and differentiable function $F \colon \mathbb{H} \to \mathbb{R}$ as follows: for all $f, g \in \mathbb{H}$,

$$\mathsf{B}_F(f \| g) = F(f) - F(g) - \langle f - g, \nabla F(g) \rangle .$$

Section E.4 presents the properties of the Bregman divergence in more detail and also contains figure E.2 which illustrates the geometric interpretation of the Bregman divergence. We generalize the definition of Bregman divergence to cover the case of convex but non-differentiable loss functions F by using the notion of *subgradient*. For a convex function $F \colon \mathbb{H} \to \mathbb{R}$, we denote by $\partial F(h)$ the *subdifferential* of F at h, which is defined as follows:

$$\partial F(h) = \{g \in \mathbb{H} \colon \forall h' \in \mathbb{H}, F(h') - F(h) \geq \langle h' - h, g \rangle\}.$$

Thus, $\partial F(h)$ is the set of vectors g defining a hyperplane supporting function F at point h (see figure 14.1). Elements of the subdifferential are called subgradients (see section B.4.1 for more discussion). Note, the subgradient found in $\partial F(h)$ coincides with $\nabla F(h)$ when F is differentiable at h, i.e. $\partial F(h) = \{\nabla F(h)\}$. Furthermore, at a point h where F is minimal, 0 is an element of $\partial F(h)$. The subgradient is additive, that is, for two convex function F_1 and F_2, $\partial(F_1 + F_2)(h) = \{g_1 + g_2 \colon g_1 \in \partial F_1(h), g_2 \in \partial F_2(h)\}$. For any $h \in \mathbb{H}$, we fix $\delta F(h)$ to be an (arbitrary) element of $\partial F(h)$. For any such choice of δF, we can define the *generalized Bregman divergence* associated to F by:

$$\forall h', h \in \mathbb{H}, \mathsf{B}_F(h' \,\|\, h) = F(h') - F(h) - \langle h' - h, \delta F(h) \rangle . \tag{14.8}$$

Note that by definition of the subgradient, $\mathsf{B}_F(h' \,\|\, h) \geq 0$ for all $h', h \in \mathbb{H}$.

Starting from (14.6), we can now define the generalized Bregman divergence of F_S. Let N denote the convex function $h \to \|h\|_K^2$. Since N is differentiable,

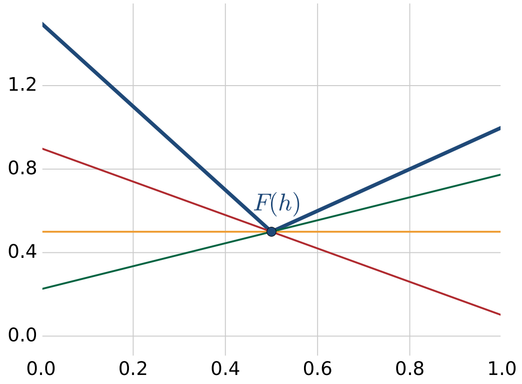

Figure 14.1
Illustration of the notion of subgradient: supporting hyperplanes (shown in red, orange and green) for the function F (shown in blue) at point h are defined by elements of the subdifferential $\partial F(h)$.

$\delta N(h) = \nabla N(h)$ for all $h \in \mathbb{H}$, and thus δN (as well as B_N) is uniquely defined. To make the definition of the Bregman divergences for F_S and $\widehat{R}_S$ compatible so that $\mathsf{B}_{F_S} = \mathsf{B}_{\widehat{R}_S} + \lambda \mathsf{B}_N$, we define $\delta \widehat{R}_S$ in terms of δF_S by: $\delta \widehat{R}_S(h) = \delta F_S(h) - \lambda \nabla N(h)$ for all $h \in \mathbb{H}$. Furthermore, we choose $\delta F_S(h)$ to be 0 for any point h where F_S is minimal and let $\delta F_S(h)$ be an arbitrary element of $\partial F_S(h)$ for all other $h \in \mathbb{H}$. We proceed in a similar way to define the Bregman divergences for $F_{S'}$ and $\widehat{R}_{S'}$ so that $\mathsf{B}_{F_{S'}} = \mathsf{B}_{\widehat{R}_{S'}} + \lambda \mathsf{B}_N$.

We will use the notion of generalized Bregman divergence for the proof of the following general upper bound on the stability coefficient of kernel-based regularization algorithms.

Proposition 14.4 *Let K be a positive definite symmetric kernel such that for all $x \in \mathcal{X}$, $K(x, x) \leq r^2$ for some $r \in \mathbb{R}_+$ and let L be a convex and σ-admissible loss function. Then, the kernel-based regularization algorithm defined by the minimization* (14.6) *is β-stable with the following upper bound on β:*

$$\beta \leq \frac{\sigma^2 r^2}{m\lambda}.$$

Proof: Let h be a minimizer of F_S and h' a minimizer of $F_{S'}$, where samples S and S' differ exactly by one point, z_m in S and z'_m in S'. Since the generalized Bregman divergence is non-negative and since $\mathsf{B}_{F_S} = \mathsf{B}_{\widehat{R}_S} + \lambda \mathsf{B}_N$ and $\mathsf{B}_{F_{S'}} = \mathsf{B}_{\widehat{R}_{S'}} + \lambda \mathsf{B}_N$, we can write

$$\mathsf{B}_{F_S}(h' \| h) + \mathsf{B}_{F_{S'}}(h \| h') \geq \lambda\big(\mathsf{B}_N(h' \| h) + \mathsf{B}_N(h \| h')\big).$$

Observe that $\mathsf{B}_N(h'\|h) + \mathsf{B}_N(h\|h') = -\langle h' - h, 2h\rangle - \langle h - h', 2h'\rangle = 2\|h' - h\|_K^2$. Let Δh denote $h' - h$, then we can write

$$\begin{aligned}
&2\lambda||\Delta h\|_K^2\\
&\quad \leq \mathsf{B}_{F_S}(h' \,\|\, h) + \mathsf{B}_{F_{S'}}(h \,\|\, h')\\
&\quad = F_S(h') - F_S(h) - \langle h' - h, \delta F_S(h)\rangle + F_{S'}(h) - F_{S'}(h') - \langle h - h', \delta F_{S'}(h')\rangle\\
&\quad = F_S(h') - F_S(h) + F_{S'}(h) - F_{S'}(h')\\
&\quad = \widehat{R}_S(h') - \widehat{R}_S(h) + \widehat{R}_{S'}(h) - \widehat{R}_{S'}(h').
\end{aligned}$$

The second equality follows from the definition of h' and h as minimizers and our choice of the subgradients for minimal points which together imply $\delta F_{S'}(h') = 0$ and $\delta F_S(h) = 0$. The last equality follows from the definitions of F_S and $F_{S'}$. Next, we express the resulting inequality in terms of the loss function L and use the fact that S and S' differ by only one point along with the σ-admissibility of L to get

$$\begin{aligned}
2\lambda\|\Delta h\|_K^2 &\leq \frac{1}{m}[L_{z_m}(h') - L_{z_m}(h) + L_{z'_m}(h) - L_{z'_m}(h')]\\
&\leq \frac{\sigma}{m}[|\Delta h(x_m)| + |\Delta h(x'_m)|]. \qquad (14.9)
\end{aligned}$$

By the reproducing kernel property and the Cauchy-Schwarz inequality, for all $x \in \mathcal{X}$,

$$\Delta h(x) = \langle \Delta h, K(x, \cdot)\rangle \leq \|\Delta h\|_K \|K(x, \cdot)\|_K = \sqrt{K(x, x)}\|\Delta h\|_K \leq r\|\Delta h\|_K.$$

In view of (14.9), this implies $\|\Delta h\|_K \leq \frac{\sigma r}{\lambda m}$. By the σ-admissibility of L and the reproducing property, the following holds:

$$\forall z \in \mathcal{X} \times \mathcal{Y}, |L_z(h') - L_z(h)| \leq \sigma|\Delta h(x)| \leq r\sigma\|\Delta h\|_K,$$

which gives

$$\forall z \in \mathcal{X} \times \mathcal{Y}, |L_z(h') - L_z(h)| \leq \frac{\sigma^2 r^2}{m\lambda},$$

and concludes the proof. □

Thus, under the assumptions of the proposition, for a fixed λ, the stability coefficient of kernel-based regularization algorithms is in $O(1/m)$.

14.3.1 Application to regression algorithms: SVR and KRR

Here, we analyze more specifically two widely used regression algorithms, Support Vector Regression (SVR) and Kernel Ridge Regression (KRR), which are both special instances of the family of kernel-based regularization algorithms.

SVR is based on the ϵ-insensitive loss L_ϵ defined for all $(y, y') \in \mathcal{Y} \times \mathcal{Y}$ by:

$$L_\epsilon(y', y) = \begin{cases} 0 & \text{if } |y' - y| \leq \epsilon; \\ |y' - y| - \epsilon & \text{otherwise.} \end{cases} \tag{14.10}$$

We now present a stability-based bound for SVR assuming that L_ϵ is bounded for the hypotheses returned by SVR (which, as we shall later see in lemma 14.7, is indeed the case when the label set $\mathcal{Y}$ is bounded).

Corollary 14.5 (Stability-based learning bound for SVR) *Assume that $K(x, x) \leq r^2$ for all $x \in \mathcal{X}$ for some $r \geq 0$ and that L_ϵ is bounded by $M \geq 0$. Let h_S denote the hypothesis returned by SVR when trained on an i.i.d. sample S of size m. Then, for any $\delta > 0$, the following inequality holds with probability at least $1 - \delta$:*

$$R(h_S) \leq \widehat{R}_S(h_S) + \frac{r^2}{m\lambda} + \Big(\frac{2r^2}{\lambda} + M\Big)\sqrt{\frac{\log \frac{1}{\delta}}{2m}}.$$

Proof: We first show that $L_\epsilon(\cdot) = L_\epsilon(\cdot, y)$ is 1-Lipschitz for any $y \in \mathcal{Y}$. For any $y', y'' \in \mathcal{Y}$, we must consider four cases. First, if $|y' - y| \leq \epsilon$ and $|y'' - y| \leq \epsilon$, then $|L_\epsilon(y'') - L_\epsilon(y')| = 0$. Second, if $|y' - y| > \epsilon$ and $|y'' - y| > \epsilon$, then $|L_\epsilon(y'') - L_\epsilon(y')| = ||y'' - y| - |y' - y|| \leq |y'' - y'|$, by the triangle inequality. Third, if $|y' - y| \leq \epsilon$ and $|y'' - y| > \epsilon$, then $|L_\epsilon(y'') - L_\epsilon(y')| = ||y'' - y| - \epsilon| = |y'' - y| - \epsilon \leq |y'' - y| - |y' - y| \leq |y'' - y'|$. Fourth, if $|y'' - y| \leq \epsilon$ and $|y' - y| > \epsilon$, by symmetry the same inequality is obtained as in the previous case.

Thus, in all cases, $|L_\epsilon(y'', y) - L_\epsilon(y', y)| \leq |y'' - y'|$. This implies in particular that L_ϵ is σ-admissible with $\sigma = 1$ for any hypothesis set $\mathcal{H}$. By proposition 14.4, under the assumptions made, SVR is β-stable with $\beta \leq \frac{r^2}{m\lambda}$. Plugging this expression into the bound of theorem 14.2 yields the result. □

We next present a stability-based bound for KRR, which is based on the square loss L_2 defined for all $y', y \in \mathcal{Y}$ by:

$$L_2(y', y) = (y' - y)^2. \tag{14.11}$$

As in the SVR setting, we assume in our analysis that L_2 is bounded for the hypotheses returned by KRR (which, as we shall later see again in lemma 14.7, is indeed the case when the label set $\mathcal{Y}$ is bounded).

Corollary 14.6 (Stability-based learning bound for KRR) *Assume that $K(x, x) \leq r^2$ for all $x \in \mathcal{X}$ for some $r \geq 0$ and that L_2 is bounded by $M \geq 0$. Let h_S denote the hypothesis returned by KRR when trained on an i.i.d. sample S of size m. Then, for any $\delta > 0$, the following inequality holds with probability at least $1 - \delta$:*

$$R(h_S) \leq \widehat{R}_S(h_S) + \frac{4Mr^2}{\lambda m} + \Big(\frac{8Mr^2}{\lambda} + M\Big)\sqrt{\frac{\log \frac{1}{\delta}}{2m}}.$$

Proof: For any $(x, y) \in \mathcal{X} \times \mathcal{Y}$ and $h, h' \in \mathcal{H}$,

$$\begin{aligned} |L_2(h'(x), y) - L_2(h(x), y)| &= \big|(h'(x) - y)^2 - (h(x) - y)^2\big| \\ &= \Big|\big[h'(x) - h(x)\big][(h'(x) - y) + (h(x) - y)]\Big| \\ &\leq (|h'(x) - y| + |h(x) - y|)|h(x) - h'(x)| \\ &\leq 2\sqrt{M}|h(x) - h'(x)|, \end{aligned}$$

where we used the M-boundedness of the loss. Thus, L_2 is σ-admissible with $\sigma = 2\sqrt{M}$. Therefore, by proposition 14.4, KRR is β-stable with $\beta \leq \frac{4r^2 M}{m\lambda}$. Plugging this expression into the bound of theorem 14.2 yields the result. □

The previous two corollaries assumed bounded loss functions. We now present a lemma that implies in particular that the loss functions used by SVR and KRR are bounded when the label set is bounded.

Lemma 14.7 *Assume that $K(x, x) \leq r^2$ for all $x \in \mathcal{X}$ for some $r \geq 0$ and that for all $y \in \mathcal{Y}$, $L(0, y) \leq B$ for some $B \geq 0$. Then, the hypothesis h_S returned by a kernel-based regularization algorithm trained on a sample S is bounded as follows:*

$$\forall x \in \mathcal{X}, |h_S(x)| \leq r\sqrt{B/\lambda}.$$

Proof: By the reproducing kernel property and the Cauchy-Schwarz inequality, we can write

$$\forall x \in \mathcal{X}, |h_S(x)| = \langle h_S, K(x, \cdot)\rangle \leq \|h_S\|_K \sqrt{K(x, x)} \leq r\|h_S\|_K. \tag{14.12}$$

The minimization (14.6) is over $\mathbb{H}$, which includes 0. Thus, by definition of F_S and h_S, the following inequality holds:

$$F_S(h_S) \leq F_S(0) = \frac{1}{m}\sum_{i=1}^{m} L(0, y_i) \leq B.$$

Since the loss L is non-negative, we have $\lambda\|h_S\|_K^2 \leq F_S(h_S)$ and thus $\lambda\|h_S\|_K^2 \leq B$. Combining this inequality with (14.12) yields the result. □

14.3.2 Application to classification algorithms: SVMs

This section presents a generalization bound for SVMs, when using the standard hinge loss defined for all $y \in \mathcal{Y} = \{-1, +1\}$ and $y' \in \mathbb{R}$ by

$$L_{\text{hinge}}(y', y) = \begin{cases} 0 & \text{if } 1 - yy' \leq 0; \\ 1 - yy' & \text{otherwise.} \end{cases} \tag{14.13}$$

Corollary 14.8 (Stability-based learning bound for SVMs) *Assume that $K(x, x) \leq r^2$ for all $x \in \mathcal{X}$ for some $r \geq 0$. Let h_S denote the hypothesis returned by SVMs when* *trained on an i.i.d. sample S of size m. Then, for any $\delta > 0$, the following inequality holds with probability at least $1 - \delta$:*

$$R(h_S) \leq \widehat{R}_S(h_S) + \frac{r^2}{m\lambda} + \Big(\frac{2r^2}{\lambda} + \frac{r}{\sqrt{\lambda}} + 1\Big)\sqrt{\frac{\log\frac{1}{\delta}}{2m}}.$$

Proof: It is straightforward to verify that $L_{\text{hinge}}(\cdot, y)$ is 1-Lipschitz for any $y \in \mathcal{Y}$ and therefore that it is σ-admissible with $\sigma = 1$. Therefore, by proposition 14.4, SVMs is β-stable with $\beta \leq \frac{r^2}{m\lambda}$. Since $|L_{\text{hinge}}(0, y)| \leq 1$ for any $y \in \mathcal{Y}$, by lemma 14.7, $\forall x \in \mathcal{X}, |h_S(x)| \leq r/\sqrt{\lambda}$. Thus, for any sample S and any $x \in \mathcal{X}$ and $y \in \mathcal{Y}$, the loss is bounded as follows: $L_{\text{hinge}}(h_S(x), y) \leq r/\sqrt{\lambda} + 1$. Plugging this value of M and the one found for β into the bound of theorem 14.2 yields the result. □

Since the hinge loss upper bounds the binary loss, the bound of the corollary 14.8 also applies to the generalization error of h_S measured in terms of the standard binary loss used in classification.

14.3.3 Discussion

Note that the learning bounds presented for kernel-based regularization algorithms are of the form $R(h_S) - \widehat{R}_S(h_S) \leq O\big(\frac{1}{\lambda\sqrt{m}}\big)$. Thus, these bounds are informative only when $\lambda \gg 1/\sqrt{m}$. The regularization parameter λ is a function of the sample size m: for larger values of m, it is expected to be smaller, decreasing the emphasis on regularization. The magnitude of λ affects the norm of the linear hypotheses used for prediction, with a larger value of λ implying a smaller hypothesis norm. In this sense, λ is a measure of the complexity of the hypothesis set and the condition required for λ can be interpreted as stating that a less complex hypothesis set guarantees better generalization.

Note also that our analysis of stability in this chapter assumed a fixed λ: the regularization parameter is assumed to be invariant to the change of one point of the training sample. While this is a mild assumption, it may not hold in general.

14.4 Chapter notes

The notion of algorithmic stability was first used by Devroye, Rogers and Wagner [Rogers and Wagner, 1978, Devroye and Wagner, 1979a,b] for the k-nearest neighbor algorithm and other k-local rules. Kearns and Ron [1999] later gave a formal definition of stability and used it to provide an analysis of the leave-one-out error. Much of the material presented in this chapter is based on Bousquet and Elisseeff [2002]. Our proof of proposition 14.4 is novel and generalizes the results of Bousquet and Elisseeff [2002] to the case of non-differentiable convex losses. Moreover, stability-based generalization bounds have been extended to ranking algorithms [Agarwal and Niyogi, 2005, Cortes et al., 2007b], as well as to the non-i.i.d.

scenario of stationary Φ- and β-mixing processes [Mohri and Rostamizadeh, 2010], and to the transductive setting [Cortes et al., 2008a]. Additionally, exercise 14.5 is based on Cortes et al. [2010b], which introduces and analyzes stability with respect to the choice of the kernel function or kernel matrix.

Note that while, as shown in this chapter, uniform stability is sufficient for deriving generalization bounds, it is not a necessary condition. Some algorithms may generalize well in the supervised learning scenario but may not be uniformly stable, for example, the Lasso algorithm [Xu et al., 2008]. Shalev-Shwartz et al. [2009] have used the notion of stability to provide necessary and sufficient conditions for a technical condition of learnability related to PAC-learning, even in general scenarios where learning is possible only by using non-ERM rules.

14.5 Exercises

14.1 Tighter stability bounds

(a) Assuming the conditions of theorem 14.2 hold, can one hope to guarantee a generalization with slack better than $O(1/\sqrt{m})$ even if the algorithm is very stable, i.e. $\beta \to 0$?

(b) Can you show an $O(1/m)$ generalization guarantee if L is bounded by $C/\sqrt{m}$ (a very strong condition)? If so, how stable does the learning algorithm need to be?

14.2 Quadratic hinge loss stability. Let L denote the quadratic hinge loss function defined for all $y \in \{+1, -1\}$ and $y' \in \mathbb{R}$ by

$$L(y', y) = \begin{cases} 0 & \text{if } 1 - y'y \leq 0; \\ (1 - y'y)^2 & \text{otherwise.} \end{cases}$$

Assume that $L(h(x), y)$ is bounded by M, $1 \leq M < \infty$, for all $h \in \mathcal{H}$, $x \in \mathcal{X}$, and $y \in \{+1, -1\}$, which also implies a bound on $|h(x)|$ for all $h \in \mathcal{H}$ and $x \in \mathcal{X}$. Derive a stability-based generalization bound for SVMs with the quadratic hinge loss.

14.3 Stability of linear regression.

(a) How does the stability bound in corollary 14.6 for ridge regression (i.e. kernel ridge regression with a linear kernel) behave as $\lambda \to 0$?

(b) Can you show a stability bound for linear regression (i.e. ridge regression with $\lambda = 0$)? If not, show a counter-example.

14.4 Kernel stability. Suppose an approximation of the kernel matrix $\mathbf{K}$, denoted $\mathbf{K}'$, is used to train the hypothesis h' (and let h denote the non-approximate hypothesis). At test time, no approximation is made, so if we let $\mathbf{k}_x = \big[K(x, x_1), \ldots, K(x, x_m)\big]^\top$ we can write $h(x) = \boldsymbol{\alpha}^\top \mathbf{k}_x$ and $h'(x) = \boldsymbol{\alpha}'^\top \mathbf{k}_x$. Show that if $\forall x, x' \in \mathcal{X}, K(x, x') \leq r$ then

$$|h'(x) - h(x)| \leq \frac{rmM}{\lambda^2} \|\mathbf{K}' - \mathbf{K}\|_2 \,.$$

(*Hint*: Use exercise 10.3)

14.5 Stability of relative-entropy regularization.

(a) Consider an algorithm that selects a distribution g over a hypothesis class which is parameterized by $\theta \in \Theta$. Given a point $z = (x, y)$ the expected loss is defined as

$$H(g, z) = \int_\Theta L(h_\theta(x), y) g(\theta) \, d\theta \,,$$

with respect to a base loss function L. Assuming the loss function L is bounded by M, show that the expected loss H is M-admissible, i.e. show $|H(g, z) - H(g', z)| \leq M \int_\Theta |g(\theta) - g'(\theta)| \, d\theta$.

(b) Consider an algorithm that minimizes the *entropy regularized* objective over the choice of distribution g:

$$F_S(g) = \underbrace{\frac{1}{m} \sum_{i=1}^m H(g, z_i)}_{\widehat{R}_S(g)} + \lambda K(g, f_0) \,.$$

Here, K is the Kullback-Leibler divergence (or relative entropy) between two distributions,

$$K(g, f_0) = \int_\Theta g(\theta) \log \frac{g(\theta)}{f_0(\theta)} \, d\theta \,, \tag{14.14}$$

and f_0 is some fixed distribution. Show that such an algorithm is stable by performing the following steps:

i. First use the fact $\frac{1}{2}(\int_\Theta |g(\theta) - g'(\theta)| \, d\theta)^2 \leq K(g, g')$ (Pinsker's inequality), to show

$$\Big(\int_\Theta |g_S(\theta) - g_{S'}(\theta)| \, d\theta\Big)^2 \leq B_{K(.,f_0)}(g \| g') + B_{K(.,f_0)}(g' \| g) \,.$$

ii. Next, let g be the minimizer of F_S and g' the minimizer of $F_{S'}$, where S and S' differ only at the index m. Show that

$$\begin{aligned} &B_{K(.,f_0)}(g\|g') + B_{K(.,f_0)}(g'\|g) \\ &\qquad\qquad \leq \frac{1}{m\lambda}\big|H(g', z_m) - H(g, z_m) + H(g, z'_m) - H(g', z'_m)\big| \\ &\qquad\qquad\qquad\qquad\qquad\qquad \leq \frac{2M}{m\lambda}\int_\Theta |g(\theta) - g'(\theta)|\, d\theta\,. \end{aligned}$$

iii. Finally, combine the results above to show that the entropy regularized algorithm is $\frac{2M^2}{m\lambda}$-stable.