13

Regularization and Stability

In the previous chapter we introduced the families of convex-Lipschitz-bounded and convex-smooth-bounded learning problems. In this section we show that all learning problems in these two families are learnable. For some learning problems of this type it is possible to show that uniform convergence holds; hence they are learnable using the ERM rule. However, this is not true for all learning problems of this type. Yet, we will introduce another learning rule and will show that it learns all convex-Lipschitz-bounded and convex-smooth-bounded learning problems.

The new learning paradigm we introduce in this chapter is called *Regularized Loss Minimization*, or RLM for short. In RLM we minimize the sum of the empirical risk and a regularization function. Intuitively, the regularization function measures the complexity of hypotheses. Indeed, one interpretation of the regularization function is the structural risk minimization paradigm we discussed in Chapter 7. Another view of regularization is as a *stabilizer* of the learning algorithm. An algorithm is considered stable if a slight change of its input does not change its output much. We will formally define the notion of stability (what we mean by "slight change of input" and by "does not change much the output") and prove its close relation to learnability. Finally, we will show that using the squared ℓ_2 norm as a regularization function stabilizes all convex-Lipschitz or convex-smooth learning problems. Hence, RLM can be used as a general learning rule for these families of learning problems.

13.1 REGULARIZED LOSS MINIMIZATION

Regularized Loss Minimization (RLM) is a learning rule in which we jointly minimize the empirical risk and a regularization function. Formally, a regularization function is a mapping $R : \mathbb{R}^d \to \mathbb{R}$, and the regularized loss minimization rule outputs a hypothesis in

$$\operatorname*{argmin}_{\mathbf{w}} \left(L_S(\mathbf{w}) + R(\mathbf{w}) \right). \tag{13.1}$$

Regularized loss minimization shares similarities with minimum description length algorithms and structural risk minimization (see Chapter 7). Intuitively, the "complexity" of hypotheses is measured by the value of the regularization function, and

the algorithm balances between low empirical risk and "simpler," or "less complex," hypotheses.

There are many possible regularization functions one can use, reflecting some prior belief about the problem (similarly to the description language in Minimum Description Length). Throughout this section we will focus on one of the most simple regularization functions: $R(\mathbf{w}) = \lambda\|\mathbf{w}\|^2$, where $\lambda > 0$ is a scalar and the norm is the ℓ_2 norm, $\|\mathbf{w}\| = \sqrt{\sum_{i=1}^{d} w_i^2}$. This yields the learning rule:

$$A(S) = \operatorname*{argmin}_{\mathbf{w}} \left(L_S(\mathbf{w}) + \lambda\|\mathbf{w}\|^2 \right). \tag{13.2}$$

This type of regularization function is often called Tikhonov regularization.

As mentioned before, one interpretation of Equation (13.2) is using structural risk minimization, where the norm of $\mathbf{w}$ is a measure of its "complexity." Recall that in the previous chapter we introduced the notion of bounded hypothesis classes. Therefore, we can define a sequence of hypothesis classes, $\mathcal{H}_1 \subset \mathcal{H}_2 \subset \mathcal{H}_3 \ldots$, where $\mathcal{H}_i = \{\mathbf{w} : \|\mathbf{w}\|_2 \leq i\}$. If the sample complexity of each $\mathcal{H}_i$ depends on i then the RLM rule is similar to the SRM rule for this sequence of nested classes.

A different interpretation of regularization is as a stabilizer. In the next section we define the notion of stability and prove that stable learning rules do not overfit. But first, let us demonstrate the RLM rule for linear regression with the squared loss.

13.1.1 Ridge Regression

Applying the RLM rule with Tikhonov regularization to linear regression with the squared loss, we obtain the following learning rule:

$$\operatorname*{argmin}_{\mathbf{w}\in\mathbb{R}^d} \left(\lambda\|\mathbf{w}\|_2^2 + \frac{1}{m}\sum_{i=1}^{m} \frac{1}{2}(\langle \mathbf{w}, \mathbf{x}_i\rangle - y_i)^2 \right). \tag{13.3}$$

Performing linear regression using Equation (13.3) is called *ridge regression*.

To solve Equation (13.3) we compare the gradient of the objective to zero and obtain the set of linear equations

$$(2\lambda m I + A)\mathbf{w} = \mathbf{b},$$

where I is the identity matrix and $A, \mathbf{b}$ are as defined in Equation (9.6), namely,

$$A = \left(\sum_{i=1}^{m} \mathbf{x}_i \mathbf{x}_i^\top \right) \quad \text{and} \quad \mathbf{b} = \sum_{i=1}^{m} y_i \mathbf{x}_i \,. \tag{13.4}$$

Since A is a positive semidefinite matrix, the matrix $2\lambda m I + A$ has all its eigenvalues bounded below by $2\lambda m$. Hence, this matrix is invertible and the solution to ridge regression becomes

$$\mathbf{w} = (2\lambda m I + A)^{-1}\mathbf{b}. \tag{13.5}$$

In the next section we formally show how regularization stabilizes the algorithm and prevents overfitting. In particular, the analysis presented in the next sections (particularly, Corollary 13.11) will yield:

Theorem 13.1. *Let $\mathcal{D}$ be a distribution over $\mathcal{X} \times [-1,1]$, where $\mathcal{X} = \{\mathbf{x} \in \mathbb{R}^d : \|\mathbf{x}\| \leq 1\}$. Let $\mathcal{H} = \{\mathbf{w} \in \mathbb{R}^d : \|\mathbf{w}\| \leq B\}$. For any $\epsilon \in (0,1)$, let $m \geq 150\, B^2/\epsilon^2$. Then, applying the ridge regression algorithm with parameter $\lambda = \epsilon/(3B^2)$ satisfies*

$$\mathop{\mathbb{E}}_{S\sim\mathcal{D}^m}[L_{\mathcal{D}}(A(S))] \;\leq\; \min_{\mathbf{w}\in\mathcal{H}} L_{\mathcal{D}}(\mathbf{w}) + \epsilon.$$

Remark 13.1. The preceding theorem tells us how many examples are needed to guarantee that the *expected value* of the risk of the learned predictor will be bounded by the approximation error of the class plus ϵ. In the usual definition of agnostic PAC learning we require that the risk of the learned predictor will be bounded with probability of at least $1-\delta$. In Exercise 13.1 we show how an algorithm with a bounded expected risk can be used to construct an agnostic PAC learner.

13.2 STABLE RULES DO NOT OVERFIT

Intuitively, a learning algorithm is stable if a small change of the input to the algorithm does not change the output of the algorithm much. Of course, there are many ways to define what we mean by "a small change of the input" and what we mean by "does not change the output much". In this section we define a specific notion of stability and prove that under this definition, stable rules do not overfit.

Let A be a learning algorithm, let $S = (z_1, \ldots, z_m)$ be a training set of m examples, and let $A(S)$ denote the output of A. The algorithm A suffers from overfitting if the difference between the true risk of its output, $L_{\mathcal{D}}(A(S))$, and the empirical risk of its output, $L_S(A(S))$, is large. As mentioned in Remark 13.1, throughout this chapter we focus on the expectation (with respect to the choice of S) of this quantity, namely, $\mathbb{E}_S[L_{\mathcal{D}}(A(S)) - L_S(A(S))]$.

We next define the notion of stability. Given the training set S and an additional example z', let $S^{(i)}$ be the training set obtained by replacing the i'th example of S with z'; namely, $S^{(i)} = (z_1, \ldots, z_{i-1}, z', z_{i+1}, \ldots, z_m)$. In our definition of stability, "a small change of the input" means that we feed A with $S^{(i)}$ instead of with S. That is, we only replace one training example. We measure the effect of this small change of the input on the output of A, by comparing the loss of the hypothesis $A(S)$ on z_i to the loss of the hypothesis $A(S^{(i)})$ on z_i. Intuitively, a good learning algorithm will have $\ell(A(S^{(i)}), z_i) - \ell(A(S), z_i) \geq 0$, since in the first term the learning algorithm does not observe the example z_i while in the second term z_i is indeed observed. If the preceding difference is very large we suspect that the learning algorithm might overfit. This is because the learning algorithm drastically changes its prediction on z_i if it observes it in the training set. This is formalized in the following theorem.

Theorem 13.2. *Let $\mathcal{D}$ be a distribution. Let $S = (z_1, \ldots, z_m)$ be an i.i.d. sequence of examples and let z' be another i.i.d. example. Let $U(m)$ be the uniform distribution over $[m]$. Then, for any learning algorithm,*

$$\mathop{\mathbb{E}}_{S\sim\mathcal{D}^m}[L_{\mathcal{D}}(A(S)) - L_S(A(S))] = \mathop{\mathbb{E}}_{(S,z')\sim\mathcal{D}^{m+1}, i\sim U(m)}[\ell(A(S^{(i)}, z_i)) - \ell(A(S), z_i)]. \quad (13.6)$$

Proof. Since S and z' are both drawn i.i.d. from $\mathcal{D}$, we have that for every i,

$$\mathop{\mathbb{E}}_{S}[L_{\mathcal{D}}(A(S))] = \mathop{\mathbb{E}}_{S,z'}[\ell(A(S), z')] = \mathop{\mathbb{E}}_{S,z'}[\ell(A(S^{(i)}), z_i)].$$

On the other hand, we can write

$$\mathop{\mathbb{E}}_{S}[L_S(A(S))] = \mathop{\mathbb{E}}_{S,i}[\ell(A(S), z_i)].$$

Combining the two equations we conclude our proof. □

When the right-hand side of Equation (13.6) is small, we say that A is a *stable* algorithm – changing a single example in the training set does not lead to a significant change. Formally,

Definition 13.3 (On-Average-Replace-One-Stable). Let $\epsilon : \mathbb{N} \to \mathbb{R}$ be a monotonically decreasing function. We say that a learning algorithm A is on-average-replace-one-stable with rate $\epsilon(m)$ if for every distribution $\mathcal{D}$

$$\mathop{\mathbb{E}}_{(S,z')\sim\mathcal{D}^{m+1}, i\sim U(m)}[\ell(A(S^{(i)}, z_i)) - \ell(A(S), z_i)] \leq \epsilon(m).$$

Theorem 13.2 tells us that a learning algorithm does not overfit if and only if it is on-average-replace-one-stable. Of course, a learning algorithm that does not overfit is not necessarily a good learning algorithm – take, for example, an algorithm A that always outputs the same hypothesis. A useful algorithm should find a hypothesis that on one hand fits the training set (i.e., has a low empirical risk) and on the other hand does not overfit. Or, in light of Theorem 13.2, the algorithm should both fit the training set and at the same time be stable. As we shall see, the parameter λ of the RLM rule balances between fitting the training set and being stable.

13.3 TIKHONOV REGULARIZATION AS A STABILIZER

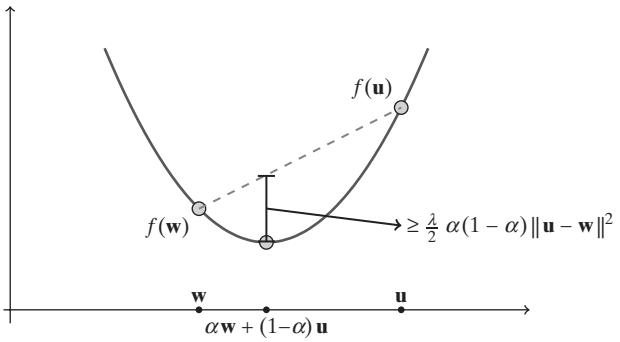
In the previous section we saw that stable rules do not overfit. In this section we show that applying the RLM rule with Tikhonov regularization, $\lambda\|\mathbf{w}\|^2$, leads to a stable algorithm. We will assume that the loss function is convex and that it is either Lipschitz or smooth.

The main property of the Tikhonov regularization that we rely on is that it makes the objective of RLM *strongly convex*, as defined in the following.

Definition 13.4 (Strongly Convex Functions). A function f is λ-strongly convex if for all $\mathbf{w}$, $\mathbf{u}$, and $\alpha \in (0,\ 1)$ we have

$$f(\alpha\mathbf{w} + (1-\alpha)\mathbf{u}) \leq \alpha f(\mathbf{w}) + (1-\alpha) f(\mathbf{u}) - \frac{\lambda}{2}\alpha(1-\alpha)\|\mathbf{w} - \mathbf{u}\|^2.$$

Clearly, every convex function is 0-strongly convex. An illustration of strong convexity is given in the following figure.



The following lemma implies that the objective of RLM is (2λ)-strongly convex. In addition, it underscores an important property of strong convexity.

Lemma 13.5.

1. *The function* $f(\mathbf{w}) = \lambda\|\mathbf{w}\|^2$ *is* 2λ*-strongly convex.*
2. *If* f *is* λ*-strongly convex and* g *is convex, then* $f + g$ *is* λ*-strongly convex.*
3. *If* f *is* λ*-strongly convex and* $\mathbf{u}$ *is a minimizer of* f*, then, for any* $\mathbf{w}$*,*

$$f(\mathbf{w}) - f(\mathbf{u}) \geq \frac{\lambda}{2}\|\mathbf{w} - \mathbf{u}\|^2.$$

Proof. The first two points follow directly from the definition. To prove the last point, we divide the definition of strong convexity by α and rearrange terms to get that

$$\frac{f(\mathbf{u} + \alpha(\mathbf{w} - \mathbf{u})) - f(\mathbf{u})}{\alpha} \leq f(\mathbf{w}) - f(\mathbf{u}) - \frac{\lambda}{2}(1 - \alpha)\|\mathbf{w} - \mathbf{u}\|^2.$$

Taking the limit $\alpha \to 0$ we obtain that the right-hand side converges to $f(\mathbf{w}) - f(\mathbf{u}) - \frac{\lambda}{2}\|\mathbf{w} - \mathbf{u}\|^2$. On the other hand, the left-hand side becomes the derivative of the function $g(\alpha) = f(\mathbf{u} + \alpha(\mathbf{w} - \mathbf{u}))$ at $\alpha = 0$. Since $\mathbf{u}$ is a minimizer of f, it follows that $\alpha = 0$ is a minimizer of g, and therefore the left-hand side of the preceding goes to zero in the limit $\alpha \to 0$, which concludes our proof. □

We now turn to prove that RLM is stable. Let $S = (z_1, \ldots, z_m)$ be a training set, let z' be an additional example, and let $S^{(i)} = (z_1, \ldots, z_{i-1}, z', z_{i+1}, \ldots, z_m)$. Let A be the RLM rule, namely,

$$A(S) = \operatorname*{argmin}_{\mathbf{w}} \left(L_S(\mathbf{w}) + \lambda\|\mathbf{w}\|^2 \right).$$

Denote $f_S(\mathbf{w}) = L_S(\mathbf{w}) + \lambda\|\mathbf{w}\|^2$, and on the basis of Lemma 13.5 we know that f_S is (2λ)-strongly convex. Relying on part 3 of the lemma, it follows that for any $\mathbf{v}$,

$$f_S(\mathbf{v}) - f_S(A(S)) \geq \lambda\|\mathbf{v} - A(S)\|^2. \tag{13.7}$$

On the other hand, for any $\mathbf{v}$ and $\mathbf{u}$, and for all i, we have

$$\begin{aligned} f_S(\mathbf{v}) - f_S(\mathbf{u}) &= L_S(\mathbf{v}) + \lambda\|\mathbf{v}\|^2 - (L_S(\mathbf{u}) + \lambda\|\mathbf{u}\|^2) \\ &= L_{S^{(i)}}(\mathbf{v}) + \lambda\|\mathbf{v}\|^2 - (L_{S^{(i)}}(\mathbf{u}) + \lambda\|\mathbf{u}\|^2) \\ &\quad + \frac{\ell(\mathbf{v}, z_i) - \ell(\mathbf{u}, z_i)}{m} + \frac{\ell(\mathbf{u}, z') - \ell(\mathbf{v}, z')}{m}. \end{aligned} \tag{13.8}$$

In particular, choosing $\mathbf{v} = A(S^{(i)})$, $\mathbf{u} = A(S)$, and using the fact that $\mathbf{v}$ minimizes $L_{S^{(i)}}(\mathbf{w}) + \lambda\|\mathbf{w}\|^2$, we obtain that

$$f_S(A(S^{(i)})) - f_S(A(S)) \leq \frac{\ell(A(S^{(i)}), z_i) - \ell(A(S), z_i)}{m} + \frac{\ell(A(S), z') - \ell(A(S^{(i)}), z')}{m}. \tag{13.9}$$

Combining this with Equation (13.7) we obtain that

$$\lambda\|A(S^{(i)}) - A(S)\|^2 \leq \frac{\ell(A(S^{(i)}), z_i) - \ell(A(S), z_i)}{m} + \frac{\ell(A(S), z') - \ell(A(S^{(i)}), z')}{m}. \tag{13.10}$$

The two subsections that follow continue the stability analysis for either Lipschitz or smooth loss functions. For both families of loss functions we show that RLM is stable and therefore it does not overfit.

13.3.1 Lipschitz Loss

If the loss function, $\ell(\cdot, z_i)$, is ρ-Lipschitz, then by the definition of Lipschitzness,

$$\ell(A(S^{(i)}), z_i) - \ell(A(S), z_i) \leq \rho\,\|A(S^{(i)}) - A(S)\|. \tag{13.11}$$

Similarly,

$$\ell(A(S), z') - \ell(A(S^{(i)}), z') \leq \rho\,\|A(S^{(i)}) - A(S)\|.$$

Plugging these inequalities into Equation (13.10) we obtain

$$\lambda\|A(S^{(i)}) - A(S)\|^2 \leq \frac{2\rho\,\|A(S^{(i)}) - A(S)\|}{m},$$

which yields

$$\|A(S^{(i)}) - A(S)\| \leq \frac{2\rho}{\lambda m}.$$

Plugging the preceding back into Equation (13.11) we conclude that

$$\ell(A(S^{(i)}), z_i) - \ell(A(S), z_i) \leq \frac{2\rho^2}{\lambda m}.$$

Since this holds for any $S,\ z',\ i$ we immediately obtain:

Corollary 13.6. *Assume that the loss function is convex and ρ-Lipschitz. Then, the RLM rule with the regularizer $\lambda\|\mathbf{w}\|^2$ is on-average-replace-one-stable with rate $\frac{2\rho^2}{\lambda m}$.*

It follows (using Theorem 13.2) that

$$\mathop{\mathbb{E}}_{S\sim\mathcal{D}^m}[L_{\mathcal{D}}(A(S)) - L_S(A(S))] \leq \frac{2\rho^2}{\lambda m}.$$

13.3.2 Smooth and Nonnegative Loss

If the loss is β-smooth and nonnegative then it is also self-bounded (see Section 12.1):

$$\|\nabla f(\mathbf{w})\|^2 \leq 2\beta f(\mathbf{w}). \tag{13.12}$$

We further assume that $\lambda \geq \frac{2\beta}{m}$, or, in other words, that $\beta \leq \lambda m/2$. By the smoothness assumption we have that

$$\ell(A(S^{(i)}), z_i) - \ell(A(S), z_i) \leq \langle \nabla \ell(A(S), z_i), A(S^{(i)}) - A(S)\rangle + \frac{\beta}{2}\|A(S^{(i)}) - A(S)\|^2. \tag{13.13}$$

Using the Cauchy-Schwartz inequality and Equation (12.6) we further obtain that

$$\begin{aligned}
&\ell(A(S^{(i)}), z_i) - \ell(A(S), z_i) \\
&\qquad \leq \|\nabla \ell(A(S), z_i)\| \, \|A(S^{(i)}) - A(S)\| + \frac{\beta}{2}\|A(S^{(i)}) - A(S)\|^2 \\
&\qquad \leq \sqrt{2\beta \ell(A(S), z_i)} \, \|A(S^{(i)}) - A(S)\| + \frac{\beta}{2}\|A(S^{(i)}) - A(S)\|^2. \qquad (13.14)
\end{aligned}$$

By a symmetric argument it holds that

$$\begin{aligned}
&\ell(A(S), z') - \ell(A(S^{(i)}), z') \\
&\qquad \leq \sqrt{2\beta \ell(A(S^{(i)}), z')} \, \|A(S^{(i)}) - A(S)\| + \frac{\beta}{2}\|A(S^{(i)}) - A(S)\|^2.
\end{aligned}$$

Plugging these inequalities into Equation (13.10) and rearranging terms we obtain that

$$\|A(S^{(i)}) - A(S)\| \leq \frac{\sqrt{2\beta}}{(\lambda m - \beta)} \left(\sqrt{\ell(A(S), z_i)} + \sqrt{\ell(A(S^{(i)}), z')} \right).$$

Combining the preceding with the assumption $\beta \leq \lambda m/2$ yields

$$\|A(S^{(i)}) - A(S)\| \leq \frac{\sqrt{8\beta}}{\lambda m} \left(\sqrt{\ell(A(S), z_i)} + \sqrt{\ell(A(S^{(i)}), z')} \right).$$

Combining the preceding with Equation (13.14) and again using the assumption $\beta \leq \lambda m/2$ yield

$$\begin{aligned}
&\ell(A(S^{(i)}), z_i) - \ell(A(S), z_i) \\
&\leq \sqrt{2\beta \ell(A(S), z_i)} \|A(S^{(i)}) - A(S)\| + \frac{\beta}{2} \|A(S^{(i)}) - A(S)\|^2 \\
&\leq \left(\frac{4\beta}{\lambda m} + \frac{8\beta^2}{(\lambda m)^2} \right) \left(\sqrt{\ell(A(S), z_i)} + \sqrt{\ell(A(S^{(i)}), z')} \right)^2 \\
&\leq \frac{8\beta}{\lambda m} \left(\sqrt{\ell(A(S), z_i)} + \sqrt{\ell(A(S^{(i)}), z')} \right)^2 \\
&\leq \frac{24\beta}{\lambda m} \left(\ell(A(S), z_i) + \ell(A(S^{(i)}), z') \right),
\end{aligned}$$

where in the last step we used the inequality $(a+b)^2 \leq 3(a^2+b^2)$. Taking expectation with respect to S, z', i and noting that $\mathbb{E}[\ell(A(S), z_i)] = \mathbb{E}[\ell(A(S^{(i)}), z')] = \mathbb{E}[L_S(A(S))]$, we conclude that:

Corollary 13.7. *Assume that the loss function is β-smooth and nonnegative. Then, the RLM rule with the regularizer $\lambda\|\mathbf{w}\|^2$, where $\lambda \geq \frac{2\beta}{m}$, satisfies*

$$\mathbb{E}\left[\ell(A(S^{(i)}), z_i) - \ell(A(S), z_i) \right] \leq \frac{48\beta}{\lambda m} \mathbb{E}[L_S(A(S))].$$

Note that if for all z we have $\ell(\mathbf{0}, z) \leq C$, for some scalar $C > 0$, then for every S,

$$L_S(A(S)) \leq L_S(A(S)) + \lambda\|A(S)\|^2 \leq L_S(\mathbf{0}) + \lambda\|\mathbf{0}\|^2 = L_S(\mathbf{0}) \leq C.$$

Hence, Corollary 13.7 also implies that

$$\mathbb{E}\left[\ell(A(S^{(i)}), z_i) - \ell(A(S), z_i) \right] \leq \frac{48\beta C}{\lambda m}.$$

13.4 CONTROLLING THE FITTING-STABILITY TRADE-OFF

We can rewrite the expected risk of a learning algorithm as

$$\mathbb{E}_S[L_{\mathcal{D}}(A(S))] = \mathbb{E}_S[L_S(A(S))] + \mathbb{E}_S[L_{\mathcal{D}}(A(S)) - L_S(A(S))]. \tag{13.15}$$

The first term reflects how well $A(S)$ fits the training set while the second term reflects the difference between the true and empirical risks of $A(S)$. As we have shown in Theorem 13.2, the second term is equivalent to the stability of A. Since our goal is to minimize the risk of the algorithm, we need that the sum of both terms will be small.

In the previous section we have bounded the stability term. We have shown that the stability term decreases as the regularization parameter, λ, increases. On the other hand, the empirical risk increases with λ. We therefore face a tradeoff between fitting and overfitting. This tradeoff is quite similar to the bias-complexity tradeoff we discussed previously in the book.

We now derive bounds on the empirical risk term for the RLM rule. Recall that the RLM rule is defined as $A(S) = \operatorname{argmin}_{\mathbf{w}} \left(L_S(\mathbf{w}) + \lambda \|\mathbf{w}\|^2 \right)$. Fix some arbitrary vector $\mathbf{w}^*$. We have

$$L_S(A(S)) \leq L_S(A(S)) + \lambda \|A(S)\|^2 \leq L_S(\mathbf{w}^*) + \lambda \|\mathbf{w}^*\|^2.$$

Taking expectation of both sides with respect to S and noting that $\mathbb{E}_S[L_S(\mathbf{w}^*)] = L_{\mathcal{D}}(\mathbf{w}^*)$, we obtain that

$$\mathop{\mathbb{E}}_{S}[L_S(A(S))] \leq L_{\mathcal{D}}(\mathbf{w}^*) + \lambda \|\mathbf{w}^*\|^2. \tag{13.16}$$

Plugging this into Equation (13.15) we obtain

$$\mathop{\mathbb{E}}_{S}[L_{\mathcal{D}}(A(S))] \leq L_{\mathcal{D}}(\mathbf{w}^*) + \lambda \|\mathbf{w}^*\|^2 + \mathop{\mathbb{E}}_{S}[L_{\mathcal{D}}(A(S)) - L_S(A(S))].$$

Combining the preceding with Corollary 13.6 we conclude:

Corollary 13.8. *Assume that the loss function is convex and ρ-Lipschitz. Then, the RLM rule with the regularization function $\lambda \|\mathbf{w}\|^2$ satisfies*

$$\forall \mathbf{w}^*, \ \mathop{\mathbb{E}}_{S}[L_{\mathcal{D}}(A(S))] \leq L_{\mathcal{D}}(\mathbf{w}^*) + \lambda \|\mathbf{w}^*\|^2 + \frac{2\rho^2}{\lambda m}.$$

This bound is often called an *oracle inequality* – if we think of $\mathbf{w}^*$ as a hypothesis with low risk, the bound tells us how many examples are needed so that $A(S)$ will be almost as good as $\mathbf{w}^*$, had we known the norm of $\mathbf{w}^*$. In practice, however, we usually do not know the norm of $\mathbf{w}^*$. We therefore usually tune λ on the basis of a validation set, as described in Chapter 11.

We can also easily derive a PAC-like guarantee[1] from Corollary 13.8 for convex-Lipschitz-bounded learning problems:

Corollary 13.9. *Let $(\mathcal{H}, Z, \ell)$ be a convex-Lipschitz-bounded learning problem with parameters ρ, B. For any training set size m, let $\lambda = \sqrt{\frac{2\rho^2}{B^2 m}}$. Then, the RLM rule with the regularization function $\lambda \|\mathbf{w}\|^2$ satisfies*

$$\mathop{\mathbb{E}}_{S}[L_{\mathcal{D}}(A(S))] \leq \min_{\mathbf{w} \in \mathcal{H}} L_{\mathcal{D}}(\mathbf{w}) + \rho B \sqrt{\frac{8}{m}}.$$

In particular, for every $\epsilon > 0$, if $m \geq \frac{8\rho^2 B^2}{\epsilon^2}$ then for every distribution $\mathcal{D}$, $\mathbb{E}_S[L_{\mathcal{D}}(A(S))] \leq \min_{\mathbf{w} \in \mathcal{H}} L_{\mathcal{D}}(\mathbf{w}) + \epsilon$.

The preceding corollary holds for Lipschitz loss functions. If instead the loss function is smooth and nonnegative, then we can combine Equation (13.16) with Corollary 13.7 to get:

Corollary 13.10. *Assume that the loss function is convex, β-smooth, and nonnegative. Then, the RLM rule with the regularization function $\lambda \|\mathbf{w}\|^2$, for $\lambda \geq \frac{2\beta}{m}$, satisfies the*

[1] Again, the bound below is on the expected risk, but using Exercise 13.1 it can be used to derive an agnostic PAC learning guarantee.

following for all $\mathbf{w}^*$*:*

$$\mathop{\mathbb{E}}_{S}[L_{\mathcal{D}}(A(S))] \leq \left(1 + \frac{48\beta}{\lambda m}\right) \mathop{\mathbb{E}}_{S}[L_S(A(S))] \leq \left(1 + \frac{48\beta}{\lambda m}\right) \left(L_{\mathcal{D}}(\mathbf{w}^*) + \lambda \|\mathbf{w}^*\|^2\right).$$

For example, if we choose $\lambda = \frac{48\beta}{m}$ we obtain from the preceding that the expected true risk of $A(S)$ is at most twice the expected empirical risk of $A(S)$. Furthermore, for this value of λ, the expected empirical risk of $A(S)$ is at most $L_{\mathcal{D}}(\mathbf{w}^*) + \frac{48\beta}{m}\|\mathbf{w}^*\|^2$.

We can also derive a learnability guarantee for convex-smooth-bounded learning problems based on Corollary 13.10.

Corollary 13.11. *Let* $(\mathcal{H}, Z, \ell)$ *be a convex-smooth-bounded learning problem with parameters* β, B*. Assume in addition that* $\ell(\mathbf{0}, z) \leq 1$ *for all* $z \in Z$*. For any* $\epsilon \in (0,1)$ *let* $m \geq \frac{150\beta B^2}{\epsilon^2}$ *and set* $\lambda = \epsilon/(3B^2)$*. Then, for every distribution* $\mathcal{D}$*,*

$$\mathop{\mathbb{E}}_{S}[L_{\mathcal{D}}(A(S))] \leq \min_{\mathbf{w} \in \mathcal{H}} L_{\mathcal{D}}(\mathbf{w}) + \epsilon.$$

13.5 SUMMARY

We introduced stability and showed that if an algorithm is stable then it does not overfit. Furthermore, for convex-Lipschitz-bounded or convex-smooth-bounded problems, the RLM rule with Tikhonov regularization leads to a stable learning algorithm. We discussed how the regularization parameter, λ, controls the tradeoff between fitting and overfitting. Finally, we have shown that all learning problems that are from the families of convex-Lipschitz-bounded and convex-smooth-bounded problems are learnable using the RLM rule. The RLM paradigm is the basis for many popular learning algorithms, including ridge regression (which we discussed in this chapter) and support vector machines (which will be discussed in Chapter 15).

In the next chapter we will present Stochastic Gradient Descent, which gives us a very practical alternative way to learn convex-Lipschitz-bounded and convex-smooth-bounded problems and can also be used for efficiently implementing the RLM rule.

13.6 BIBLIOGRAPHIC REMARKS

Stability is widely used in many mathematical contexts. For example, the necessity of stability for so-called inverse problems to be well posed was first recognized by Hadamard (1902). The idea of regularization and its relation to stability became widely known through the works of Tikhonov (1943) and Phillips (1962). In the context of modern learning theory, the use of stability can be traced back at least to the work of Rogers and Wager (1978), which noted that the sensitivity of a learning algorithm with regard to small changes in the sample controls the variance of the leave-one-out estimate. The authors used this observation to obtain generalization bounds for the k-nearest neighbor algorithm (see Chapter 19). These results were later extended to other "local" learning algorithms (see Devroye, Györfi & Lugosi (1996) and references therein). In addition, practical methods have been developed

to introduce stability into learning algorithms, in particular the Bagging technique introduced by (Breiman 1996).

Over the last decade, stability was studied as a generic condition for learnability. See (Kearns & Ron 1999, Bousquet & Elisseeff 2002, Kutin & Niyogi 2002, Rakhlin, Mukherjee & Poggio 2005, Mukherjee, Niyogi, Poggio & Rifkin 2006). Our presentation follows the work of Shalev-Shwartz, Shamir, Srebro, and Sridharan (2010), who showed that stability is sufficient and necessary for learning. They have also shown that all convex-Lipschitz-bounded learning problems are learnable using RLM, even though for some convex-Lipschitz-bounded learning problems uniform convergence does not hold in a strong sense.

13.7 EXERCISES

13.1 **From Bounded Expected Risk to Agnostic PAC Learning:** Let A be an algorithm that guarantees the following: If $m \geq m_{\mathcal{H}}(\epsilon)$ then for every distribution $\mathcal{D}$ it holds that

$$\mathop{\mathbb{E}}_{S\sim\mathcal{D}^m}[L_{\mathcal{D}}(A(S))] \leq \min_{h\in\mathcal{H}} L_{\mathcal{D}}(h) + \epsilon.$$

- Show that for every $\delta \in (0,1)$, if $m \geq m_{\mathcal{H}}(\epsilon\,\delta)$ then with probability of at least $1-\delta$ it holds that $L_{\mathcal{D}}(A(S)) \leq \min_{h\in\mathcal{H}} L_{\mathcal{D}}(h) + \epsilon$.
  Hint: Observe that the random variable $L_{\mathcal{D}}(A(S)) - \min_{h\in\mathcal{H}} L_{\mathcal{D}}(h)$ is nonnegative and rely on Markov's inequality.
- For every $\delta \in (0,1)$ let

$$m_{\mathcal{H}}(\epsilon,\ \delta) = m_{\mathcal{H}}(\epsilon/2)\lceil\log_2(1/\delta)\rceil + \left\lceil \frac{\log(4/\delta) + \log(\lceil\log_2(1/\delta)\rceil)}{\epsilon^2} \right\rceil.$$

  Suggest a procedure that agnostic PAC learns the problem with sample complexity of $m_{\mathcal{H}}(\epsilon,\delta)$, assuming that the loss function is bounded by 1.
  Hint: Let $k = \lceil\log_2(1/\delta)\rceil$. Divide the data into $k+1$ chunks, where each of the first k chunks is of size $m_{\mathcal{H}}(\epsilon/2)$ examples. Train the first k chunks using A. On the basis of the previous question argue that the probability that for all of these chunks we have $L_{\mathcal{D}}(A(S)) > \min_{h\in\mathcal{H}} L_{\mathcal{D}}(h) + \epsilon$ is at most $2^{-k} \leq \delta/2$. Finally, use the last chunk as a validation set.

13.2 **Learnability without Uniform Convergence:** Let $\mathcal{B}$ be the unit ball of $\mathbb{R}^d$, let $\mathcal{H} = \mathcal{B}$, let $Z = \mathcal{B} \times \{0,1\}^d$, and let $\ell : Z \times \mathcal{H} \to \mathbb{R}$ be defined as follows:

$$\ell(\mathbf{w}, (\mathbf{x},\ \boldsymbol{\alpha})) = \sum_{i=1}^{d} \alpha_i (x_i - w_i)^2.$$

This problem corresponds to an *unsupervised* learning task, meaning that we do not try to predict the label of $\mathbf{x}$. Instead, what we try to do is to find the "center of mass" of the distribution over $\mathcal{B}$. However, there is a twist, modeled by the vectors $\boldsymbol{\alpha}$. Each example is a pair $(\mathbf{x}, \boldsymbol{\alpha})$, where $\mathbf{x}$ is the instance $\mathbf{x}$ and $\boldsymbol{\alpha}$ indicates which features of $\mathbf{x}$ are "active" and which are "turned off." A hypothesis is a vector $\mathbf{w}$ representing the center of mass of the distribution, and the loss function is the squared Euclidean distance between $\mathbf{x}$ and $\mathbf{w}$, but only with respect to the "active" elements of $\mathbf{x}$.

- Show that this problem is learnable using the RLM rule with a sample complexity that does not depend on d.

- Consider a distribution $\mathcal{D}$ over Z as follows: $\mathbf{x}$ is fixed to be some $\mathbf{x}_0$, and each element of $\boldsymbol{\alpha}$ is sampled to be either 1 or 0 with equal probability. Show that the rate of uniform convergence of this problem grows with d.
  Hint: Let m be a training set size. Show that if $d \gg 2^m$, then there is a high probability of sampling a set of examples such that there exists some $j \in [d]$ for which $\alpha_j = 1$ for all the examples in the training set. Show that such a sample cannot be ϵ-representative. Conclude that the sample complexity of uniform convergence must grow with $\log(d)$.
- Conclude that if we take d to infinity we obtain a problem that is learnable but for which the uniform convergence property does not hold. Compare to the fundamental theorem of statistical learning.

13.3 **Stability and Asymptotic ERM Are Sufficient for Learnability:**
We say that a learning rule A is an *AERM (Asymptotic Empirical Risk Minimizer)* with rate $\epsilon(m)$ if for every distribution $\mathcal{D}$ it holds that

$$\mathop{\mathbb{E}}_{S\sim\mathcal{D}^m}\left[L_S(A(S)) - \min_{h\in\mathcal{H}} L_S(h)\right] \leq \epsilon(m).$$

We say that a learning rule A learns a class $\mathcal{H}$ with rate $\epsilon(m)$ if for every distribution $\mathcal{D}$ it holds that

$$\mathop{\mathbb{E}}_{S\sim\mathcal{D}^m}\left[L_{\mathcal{D}}(A(S)) - \min_{h\in\mathcal{H}} L_{\mathcal{D}}(h)\right] \leq \epsilon(m).$$

Prove the following:

Theorem 13.12. *If a learning algorithm A is on-average-replace-one-stable with rate $\epsilon_1(m)$ and is an AERM with rate $\epsilon_2(m)$, then it learns $\mathcal{H}$ with rate $\epsilon_1(m) + \epsilon_2(m)$.*

13.4 **Strong Convexity with Respect to General Norms:**
Throughout the section we used the ℓ_2 norm. In this exercise we generalize some of the results to general norms. Let $\|\cdot\|$ be some arbitrary norm, and let f be a strongly convex function with respect to this norm (see Definition 13.4).

1. Show that items 2–3 of Lemma 13.5 hold for every norm.
2. **(*)** Give an example of a norm for which item 1 of Lemma 13.5 does not hold.
3. Let $R(\mathbf{w})$ be a function that is (2λ)-strongly convex with respect to some norm $\|\cdot\|$. Let A be an RLM rule with respect to R, namely,

$$A(S) = \operatorname*{argmin}_{\mathbf{w}} \left(L_S(\mathbf{w}) + R(\mathbf{w})\right).$$

   Assume that for every z, the loss function $\ell(\cdot, z)$ is ρ-Lipschitz with respect to the same norm, namely,

$$\forall z,\ \forall \mathbf{w},\ \mathbf{v},\quad \ell(\mathbf{w},\ z) - \ell(\mathbf{v},\ z) \leq \rho\,\|\mathbf{w} - \mathbf{v}\|.$$

   Prove that A is on-average-replace-one-stable with rate $\frac{2\rho^2}{\lambda m}$.
4. **(*)** Let $q \in (1,\ 2)$ and consider the ℓ_q-norm

$$\|\mathbf{w}\|_q = \left(\sum_{i=1}^{d} |w_i|^q\right)^{1/q}.$$

It can be shown (see, for example, Shalev-Shwartz (2007)) that the function

$$R(\mathbf{w}) = \frac{1}{2(q-1)}\|\mathbf{w}\|_q^2$$

is 1-strongly convex with respect to $\|\mathbf{w}\|_q$. Show that if $q = \frac{\log(d)}{\log(d)-1}$ then $R(\mathbf{w})$ is $\left(\frac{1}{3\log(d)}\right)$-strongly convex with respect to the ℓ_1 norm over $\mathbb{R}^d$.