6

Basic Statistical Analysis of SVMs

Overview. *So far we have not considered the fact that SVMs typically deal with observations from a random process. This chapter addresses this issue by so-called oracle inequalities that relate the risk of an empirical SVM solution to that of the corresponding infinite-sample SVM. In particular, we will see that the analysis of the learning ability of SVMs can be split into a statistical part described by the oracle inequalities and a deterministic part based on the approximation error function investigated in the previous chapter.*

Prerequisites. *The first three sections require only basic knowledge of probability theory as well as some notions from the introduction in Chapter 1 and Sections 2.1 and 2.2 on loss functions. In the last two sections, we additionally need Sections 4.2, 4.3, and 4.6 on kernels and Chapter 5 on infinite-sample SVMs.*

Usage. *The oracle inequalities of this chapter are necessary for Chapter 8 on classification. In addition, they are helpful in Chapter 11, where practical strategies for selecting hyper parameters are discussed.*

Let us recall from the introduction that the goal of learning from a training set D is to find a decision function f_D such that $\mathcal{R}_{L,\mathrm{P}}(f_D)$ is close to the minimal risk $\mathcal{R}^*_{L,\mathrm{P}}$. Since we typically assume that the empirical data set D consists of i.i.d. observations from an unknown distribution P, the decision function f_D and its associated risk $\mathcal{R}_{L,\mathrm{P}}(f_D)$ become random variables. Informally, the "learning ability" of a learning method $D \mapsto f_D$ can hence be described by an answer to the following question:

What is the probability that $\mathcal{R}_{L,\mathrm{P}}(f_D)$ is close to $\mathcal{R}^*_{L,\mathrm{P}}$?

The main goal of this chapter is to present basic concepts and techniques for addressing this question for SVMs. To this end, we introduce two key notions of statistical learning in Section 6.1, namely *consistency* and *learning rates*, that formalize possible answers to the question above. While consistency is of purely asymptotic nature, learning rates provide a framework that is more closely related to practical needs. On the other hand, we will see in Section 6.1 that consistency can often be ensured *without* assumptions on P, while learning with guaranteed rates *almost always* requires assumptions on the unknown distribution P. In the second section, we then establish some basic concentration inequalities that will be the key tools for investigating the statistical properties of SVMs in this chapter. Subsequently we will illustrate their use in Section 6.3, where we investigate empirical risk minimization. The last two sections are devoted to the actual statistical analysis of SVMs. In Section 6.4, we establish two *oracle inequalities* that, for a fixed regularization parameter, relate the risk of empirical SVM decision functions to the approximation error function. These oracle inequalities will then be used to establish basic forms of both consistency and learning rates for SVMs using *a priori* defined regularization parameters. Thereby it turns out that the fastest learning rates our analysis provides require some knowledge about the distribution P. Unfortunately, however, the required type of knowledge on P is rarely available in practice, and hence these rates are in general not achievable with *a priori* defined regularization parameters. Finally, in Section 6.5 we present and analyze a simple method for determining the regularization parameter for SVMs in a *data-dependent* way. Here it will turn out that this method is *adaptive* in the sense that it achieves the fastest learning rates of our previous analysis *without* knowing any characteristics of P.

6.1 Notions of Statistical Learning

In this section, we introduce some basic notions that describe "learning" in a more formal sense. Let us begin by defining learning methods.

Definition 6.1. *Let X be a set and $Y \subset \mathbb{R}$. A* ***learning method*** $\mathcal{L}$ *on $X \times Y$ maps every data set $D \in (X \times Y)^n$, $n \geq 1$, to a function $f_D : X \to \mathbb{R}$.*

By definition, *any* method that assigns to every training set D of arbitrary but finite length a function f_D is a learning method. In particular, the meaning of "learning" is not specified in this definition. However, before we can define what we actually mean by "learning", we have to introduce a rather technical assumption dealing with the measurability of learning methods. Fortunately, we will see later that this measurability is usually fulfilled for SVMs and related learning methods. Therefore, readers not interested in these technical, yet mathematically important, details may jump directly to Definition 6.4.

Before we introduce the required measurability notion for learning methods, let us first recall (see Lemma A.3.3) that the P-completion $\mathcal{A}_\mathrm{P}$ of a σ-algebra $\mathcal{A}$ is the smallest σ-algebra that contains $\mathcal{A}$ and all subsets of P-zero sets in $\mathcal{A}$. Moreover, the universal completion of $\mathcal{A}$ is defined as the intersection of all completions $\mathcal{A}_\mathrm{P}$, where P runs through the set of all probability measures defined on $\mathcal{A}$. In order to avoid notational overload, we *always* assume in this chapter that $(X \times Y)^n$ is equipped with the universal completion of the product σ-algebra on $(X \times Y)^n$, where the latter is usually defined from the context. Moreover, the canonical extension of a product measure P^n to this completion will also be denoted by P^n if no confusion can arise.

Definition 6.2. *Let $X \neq \emptyset$ be a set equipped with some σ-algebra and $Y \subset \mathbb{R}$ be a closed non-empty subset equipped with the Borel σ-algebra. We say that the* ***learning method*** *$\mathcal{L}$ on $X \times Y$ is* ***measurable*** *if for all $n \geq 1$ the map*

$$\begin{aligned} (X \times Y)^n \times X &\to \mathbb{R} \\ (D, x) &\mapsto f_D(x) \end{aligned}$$

is measurable with respect to the universal completion of the product σ-algebra on $(X \times Y)^n \times X$, where f_D denotes the decision function produced by $\mathcal{L}$.

In the following sections, we will see that both ERM and SVMs are measurable learning methods under rather natural assumptions, and therefore we omit presenting examples of measurable learning methods in this section.

Now note that for measurable learning methods the maps $x \mapsto f_D(x)$ are measurable for all fixed $D \in (X \times Y)^n$. Consequently, the risks $\mathcal{R}_{L,\mathrm{P}}(f_D)$ are defined for all $D \in (X \times Y)^n$ and all $n \geq 1$. The following lemma ensures that for measurable learning methods the "probability" for sets of the form $\{D \in (X \times Y)^n : \mathcal{R}_{L,\mathrm{P}}(f_D) \leq \varepsilon\}$, $\varepsilon \geq 0$, is defined.

Lemma 6.3. *Let $\mathcal{L}$ be a measurable learning method on $X \times Y$ producing the decision functions f_D. Then, for all loss functions $L : X \times Y \times \mathbb{R} \to [0, \infty)$, all probability measures* P *on $X \times Y$, and all $n \geq 1$, the maps*

$$\begin{aligned} (X \times Y)^n &\to [0, \infty] \\ D &\mapsto \mathcal{R}_{L,\mathrm{P}}(f_D) \end{aligned}$$

are measurable with respect to the universal completion of the product σ-algebra on $(X \times Y)^n$.

Proof. By the measurability of $\mathcal{L}$ and L, we obtain the measurability of the map $(D, x, y) \mapsto L(x, y, f_D(x))$. Now the assertion follows from the measurability statement in Tonelli's Theorem A.3.10. □

With these preparations, we can now introduce our first notion of learning.

Definition 6.4. *Let $L : X \times Y \times \mathbb{R} \to [0, \infty)$ be a loss,* P *be a distribution on $X \times Y$, and $\mathcal{L}$ be a measurable learning method on $X \times Y$. Then $\mathcal{L}$ is said to be* ***L-risk consistent*** *for* P *if, for all $\varepsilon > 0$, we have*

$$\lim_{n \to \infty} \mathrm{P}^n\Big(D \in (X \times Y)^n : \mathcal{R}_{L,\mathrm{P}}(f_D) \leq \mathcal{R}^*_{L,\mathrm{P}} + \varepsilon\Big) = 0\,. \qquad (6.1)$$

Moreover, $\mathcal{L}$ is called ***universally L-risk consistent*** *if it is L-risk consistent for all distributions* P *on $X \times Y$.*

When the training set is sufficiently large, consistent learning methods produce nearly optimal decision functions with high probability. In other words, in the long run, a consistent method is able to "learn" with high probability decision functions that achieve nearly optimally the learning goal defined by the loss function L. Moreover, universally consistent learning methods accomplish this *without* knowing any specifics of the data-generating distribution P. From an orthodox machine learning point of view, which prohibits assumptions on P, universal consistency is thus a minimal requirement for any reasonable learning method. However, one might wonder whether this point of view, though mathematically compelling, is, at least sometimes, too unrealistic. To illustrate this suspicion, let us consider binary classification on $X := [0, 1]$. Then every Bayes decision function is of the form $\mathbf{1}_{X_1} - \mathbf{1}_{X_{-1}}$, where $X_{-1}, X_1 \subset [0, 1]$ are suitable sets. In addition, let us restrict our discussion to nontrivial distributions P on $X \times \{-1, 1\}$, i.e., to distributions satisfying both $\mathrm{P}_X(X_1) > 0$ and $\mathrm{P}_X(X_{-1}) > 0$. For "elementary" classification problems, where X_1 and X_{-1} are *finite* unions of intervals, consistency then seems to be a natural minimal requirement for any reasonable learning methods. On the other hand, "monster" distributions P, such as the one where X_1 is the Cantor set, X_{-1} is its complement, and P_X is a mixture of the Hausdorff measure on X_1 and the Lebesgue measure on $[-1, 1]$, seem to be less realistic for practical applications, and hence it may be harder to argue that learning methods should be able to learn for such P. In many situations, however, we cannot *a priori* exclude elementary distributions that are disturbed by some small yet not vanishing amount attributed to a malign or even monster distribution. Therefore, universal consistency can also be viewed as a notion of robustness that prevents a learning method from asymptotically failing in the presence of deviations from (implicitly) assumed features of P.

One of the drawbacks of the notion of (universal) consistency is that it does not specify the *speed* of convergence in (6.1). In other words, consistency is a truly asymptotic notion in the sense that it does not give us any confidence about how well the method has learned for a given data set D of fixed length n. Therefore, our next goal is to introduce a notion of learning that has a less asymptotic nature. We begin by reformulating consistency.

Lemma 6.5 (Learning rate). *Let $L : X \times Y \times \mathbb{R} \to [0, \infty)$ be a loss, P be a distribution on $X \times Y$, and $\mathcal{L}$ be a measurable learning method satisfying*

$$\sup_{D \in (X \times Y)^n} \mathcal{R}_{L,\mathrm{P}}(f_D) < \infty\,, \qquad\qquad n \geq 1,$$

for its decision functions f_D. Then the following statements are equivalent:

i) $\mathcal{L}$ is L-risk consistent for P.

ii) There exist a constant $c_\mathrm{P} > 0$ and a decreasing sequence $(\varepsilon_n) \subset (0, 1]$ that converges to 0 such that for all $\tau \in (0, 1]$ there exists a constant $c_\tau \in [1, \infty)$ only depending on τ such that, for all $n \geq 1$ and all $\tau \in (0, 1]$, we have

$$\mathrm{P}^n\Big(D \in (X \times Y)^n : \mathcal{R}_{L,\mathrm{P}}(f_D) \leq \mathcal{R}^*_{L,\mathrm{P}} + c_\mathrm{P}\, c_\tau\, \varepsilon_n\Big) \;\geq\; 1 - \tau\,. \qquad (6.2)$$

In this case, $\mathcal{L}$ is said to ***learn with rate*** *(ε_n)* ***and confidence*** *$(c_\tau)_{\tau \in (0,1]}$.*

Proof. The fact that (6.2) implies L-risk consistency is trivial, and hence it remains to show the converse implication. To this end, we define the function $F : (0,\infty) \times \mathbb{N} \to [0,\infty)$ by

$$F(\varepsilon, n) := \mathrm{P}^n\Big(D \in (X\times Y)^n : \mathcal{R}_{L,\mathrm{P}}(f_D) > \mathcal{R}^*_{L,\mathrm{P}} + \varepsilon\Big)\,, \qquad \varepsilon \in (0,\infty),\, n \geq 1.$$

The L-risk consistency then shows $\lim_{n\to\infty} F(\varepsilon, n) = 0$ for all $\varepsilon > 0$, and hence Lemma A.1.4 yields a decreasing sequence $(\varepsilon_n) \subset (0,1]$ converging to 0 such that $\lim_{n\to\infty} F(\varepsilon_n, n) = 0$. For a fixed $\tau \in (0,1]$, there consequently exists an $n_\tau \geq 1$ such that

$$\mathrm{P}^n\Big(D \in (X \times Y)^n : \mathcal{R}_{L,\mathrm{P}}(f_D) > \mathcal{R}^*_{L,\mathrm{P}} + \varepsilon_n\Big) \leq \tau\,, \qquad n > n_\tau. \tag{6.3}$$

For $n \geq 1$, we write $b_n := \sup_{D\in(X\times Y)^n} \mathcal{R}_{L,\mathrm{P}}(f_D) - \mathcal{R}^*_{L,\mathrm{P}}$. Then the boundedness of $\mathcal{L}$ shows $b_n < \infty$ for all $n \geq 1$ and, by the definition of b_n, we also have

$$\mathrm{P}^n\Big(D \in (X \times Y)^n : \mathcal{R}_{L,\mathrm{P}}(f_D) > \mathcal{R}^*_{L,\mathrm{P}} + b_n\Big) \leq \tau\,, \qquad n = 1, \dots, n_\tau. \tag{6.4}$$

Let us define $c_\tau := \varepsilon_{n_\tau}^{-1} \max\{1, b_1, \dots, b_{n_\tau}\}$. Then we have $b_n \leq c_\tau \varepsilon_n$ for all $n = 1, \dots, n_\tau$ and, since $c_\tau \geq 1$, we also have $\varepsilon_n \leq c_\tau \varepsilon_n$ for all $n > n_\tau$. Using these estimates in (6.4) and (6.3), respectively, yields (6.2). □

Note that in (6.2) the constant c_P depends on the distribution P. Therefore, if we do not know P, then in general we do not know c_P. In other words, even if we know that $\mathcal{L}$ learns with rate (ε_n) for all distributions P, this knowledge does not give us any confidence about how well the method has learned in a specific application. Unfortunately, however, the following results show that the situation is even worse in the sense that in general there exists no method that learns with a fixed rate and confidence for all distributions P. Before we state these results, we remind the reader that Lebesgue absolutely continuous distributions on subsets of $\mathbb{R}^d$ are atom-free. A precise definition of atom-free measures is given in Definition A.3.12.

Theorem 6.6 (No-free-lunch theorem). *Let $(a_n) \subset (0, 1/16]$ be a decreasing sequence that converges to 0. Moreover, let $(X, \mathcal{A}, \mu)$ be an atom-free probability space, $Y := \{-1, 1\}$, and L_class be the binary classification loss. Then, for every measurable learning method $\mathcal{L}$ on $X \times Y$, there exists a distribution P on $X \times Y$ with $\mathrm{P}_X = \mu$ such that $\mathcal{R}^*_{L_\mathrm{class},\mathrm{P}} = 0$ and*

$$\mathbb{E}_{D\sim\mathrm{P}^n} \mathcal{R}_{L_\mathrm{class},\mathrm{P}}(f_D) \geq a_n\,, \qquad n \geq 1.$$

Since the proof of this theorem is out of the scope of this book, we refer the interested reader to Theorem 7.2 in the book by Devroye *et al.* (1996). In this regard, we also note that Lyapunov's Theorem A.3.13 can be easily utilized to generalize their proof to a fixed atom-free distribution. The details are discussed in Exercise 6.4.

Informally speaking, the no-free-lunch theorem states that, for sufficiently malign distributions, the *average* risk of any classification method may tend arbitrarily slowly to zero. Our next goal is to use this theorem to show that in general no learning method enjoys a uniform learning rate. The first result in this direction deals with the classification loss.

Corollary 6.7 (No uniform rate for classification). *Let $(X, \mathcal{A}, \mu)$ be an atom-free probability space, $Y := \{-1, 1\}$, and $\mathcal{L}$ be a measurable learning method on $X \times Y$. Then, for all decreasing sequences $(\varepsilon_n) \subset (0, 1]$ that converge to 0 and all families $(c_\tau)_{t\in(0,1]} \subset [1, \infty)$, there exists a distribution* P *on $X \times Y$ satisfying $\mathrm{P}_X = \mu$ and $\mathcal{R}^*_{L_{\text{class}},\mathrm{P}} = 0$ such that $\mathcal{L}$ does not learn with rate (ε_n) and confidence $(c_\tau)_{\tau\in(0,1]}$.*

Proof. For brevity's sake, we write $L := L_{\text{class}}$. Let us assume that the assertion is false, i.e., that there exist a decreasing sequence $(\varepsilon_n) \subset (0, 1]$ that converges to 0 and constants $c_\tau \in [1, \infty)$, $\tau \in (0, 1]$, such that, for all distributions P on $X \times Y$ satisfying $\mathrm{P}_X = \mu$ and $\mathcal{R}^*_{L,\mathrm{P}} = 0$, the method $\mathcal{L}$ learns with rate (ε_n) and confidence $(c_\tau)_{\tau\in(0,1]}$. In other words, we assume

$$\mathrm{P}^n\Big(D \in (X \times Y)^n : \mathcal{R}_{L,\mathrm{P}}(f_D) > c_\mathrm{P}\, c_\tau\, \varepsilon_n\Big) \;\leq\; \tau \tag{6.5}$$

for all $n \geq 1$ and $\tau \in (0, 1]$, where c_P is a constant independent of n and τ. Let us define $F : (0, 1] \times \mathbb{N} \to [0, \infty)$ by $F(\tau, n) := \tau^{-1} c_\tau \varepsilon_n$. Then we have $\lim_{n\to\infty} F(\tau, n) = 0$ for all $\tau \in (0, 1]$, and consequently an obvious modification of Lemma A.1.4 yields a decreasing sequence $(\tau_n) \subset (0, 1]$ converging to 0 such that $\lim_{n\to\infty} F(\tau_n, n) = 0$. We define $a_n := 1/16$ if $\tau_n \geq 1/32$ and $a_n := 2\tau_n$ otherwise. By the no-free-lunch theorem, there then exists a distribution P on $X \times Y$ such that $\mathrm{P}_X = \mu$, $\mathcal{R}^*_{L,\mathrm{P}} = 0$, and

$$\begin{aligned} a_n &\leq \mathbb{E}_{D\sim \mathrm{P}^n} \mathcal{R}_{L,\mathrm{P}}(f_D) \\ &= \int\limits_{\mathcal{R}_{L,\mathrm{P}}(f_D)\leq c_\mathrm{P}\, c_\tau\, \varepsilon_n} \mathcal{R}_{L,\mathrm{P}}(f_D)\, d\mathrm{P}^n(D) + \int\limits_{\mathcal{R}_{L,\mathrm{P}}(f_D)> c_\mathrm{P}\, c_\tau\, \varepsilon_n} \mathcal{R}_{L,\mathrm{P}}(f_D)\, d\mathrm{P}^n(D) \\ &\leq c_\mathrm{P}\, c_\tau\, \varepsilon_n + \tau \end{aligned}$$

for all $n \geq 1$ and $\tau \in (0, 1]$, where in the last estimate we used (6.5) together with $\mathcal{R}_{L,\mathrm{P}}(f_D) \leq 1$. Consequently, we find

$$\frac{a_n - \tau_n}{c_{\tau_n} \varepsilon_n} \;\leq\; c_\mathrm{P}\,, \qquad\qquad n \geq 1.$$

On the other hand, our construction yields

$$\lim_{n\to\infty} \frac{a_n - \tau_n}{c_{\tau_n} \varepsilon_n} = \lim_{n\to\infty} \frac{\tau_n}{c_{\tau_n} \varepsilon_n} = \lim_{n\to\infty} \frac{1}{F(\tau_n, n)} = \infty\,,$$

and hence we have found a contradiction. □

In the following, we show that the result of Corollary 6.7 is true not only for the classification loss but for basically all loss functions. The idea of this generalization is that whenever we have a loss function L that describes a learning goal where at least two different labels have to be distinguished, then this learning problem is in some sense harder than binary classification and hence cannot be learned with a uniform rate. Since the precise statement of this idea is rather cumbersome and requires notations from Section 3.1, we suggest that the first-time reader skips this part and simply remembers the informal result described above. Moreover, for *convex* losses L, the conditions below can be substantially simplified. We refer the reader to Exercise 6.5 for a precise statement and examples.

Corollary 6.8 (No uniform learning rate). *Let $(X, \mathcal{A}, \mu)$ be an atom-free probability space, $Y \subset \mathbb{R}$ be a closed subset, and $L : X \times Y \times \mathbb{R} \to [0, \infty)$ be a loss. Assume that there exist two distributions Q_1 and Q_2 on Y such that $\mathcal{M}_{L,\mathrm{Q}_1,x}(0^+) \neq \emptyset$, $\mathcal{M}_{L,\mathrm{Q}_2,x}(0^+) \neq \emptyset$, and*

$$\overline{\mathcal{M}_{L,\mathrm{Q}_1,x}(0^+)} \cap \mathcal{M}_{L,\mathrm{Q}_2,x}(0^+) = \emptyset$$

for all $x \in X$. For $x \in X$, we define

$$\mathcal{M}_{1,x} := \bigl\{t \in \mathbb{R} : \operatorname{dist}(t, \mathcal{M}_{L,\mathrm{Q}_2,x}(0^+)) \geq \operatorname{dist}(t, \mathcal{M}_{L,\mathrm{Q}_1,x}(0^+))\bigr\} ,$$
$$\mathcal{M}_{2,x} := \bigl\{t \in \mathbb{R} : \operatorname{dist}(t, \mathcal{M}_{L,\mathrm{Q}_2,x}(0^+)) < \operatorname{dist}(t, \mathcal{M}_{L,\mathrm{Q}_1,x}(0^+))\bigr\} .$$

If there exists a measurable $h : X \to (0, 1]$ such that for all $x \in X$ we have

$$\inf_{t \in \mathcal{M}_{1,x}} \mathcal{C}_{L,\mathrm{Q}_2,x}(t) - \mathcal{C}^*_{L,\mathrm{Q}_2,x} \geq h(x) ,$$
$$\inf_{t \in \mathcal{M}_{2,x}} \mathcal{C}_{L,\mathrm{Q}_1,x}(t) - \mathcal{C}^*_{L,\mathrm{Q}_1,x} \geq h(x) ,$$

then the conclusion of Corollary 6.7 remains true if we replace L_{class} by L.

Proof. Clearly, the distribution $\bar{\mu} := \|h\|^{-1}_{L_1(\mu)} h\mu$ on X is atom-free. Moreover, for a distribution P on $X \times Y$ that is of type $\{\mathrm{Q}_1, \mathrm{Q}_2\}$ and satisfies $\mathrm{P}_X = \mu$, we associate the distribution $\bar{\mathrm{P}}$ on $X \times \{-1, 1\}$ that is defined by $\bar{\mathrm{P}}_X = \bar{\mu}$ and

$$\bar{\mathrm{P}}(y = 1|x) := \begin{cases} 0 & \text{if } \mathrm{P}(\,\cdot\,|x) = \mathrm{Q}_1 \\ 1 & \text{if } \mathrm{P}(\,\cdot\,|x) = \mathrm{Q}_2 \,. \end{cases}$$

In other words, $\bar{\mathrm{P}}(\,\cdot\,|x)$ produces almost surely a negative label if $\mathrm{P}(\,\cdot\,|x) = \mathrm{Q}_1$ and almost surely a positive label if $\mathrm{P}(\,\cdot\,|x) = \mathrm{Q}_2$. From this it becomes obvious that $\mathcal{R}^*_{L_{\mathrm{class}},\bar{\mathrm{P}}} = 0$. For $x \in X$ and $t \in \mathbb{R}$, we further define

$$\pi(t, x) := \begin{cases} -1 & \text{if } t \in \mathcal{M}_{1,x} \\ 1 & \text{if } t \in \mathcal{M}_{2,x} \,. \end{cases}$$

In other words, $\pi(t,x)$ becomes negative if t is closer to the minimizing set $\mathcal{M}_{L,\mathrm{Q}_1,x}(0^+)$ of the distribution Q_1 than to that of Q_2. Here the idea behind this construction is that Q_1 is identified with negative classification labels in the definition of $\bar{\mathrm{P}}$. Moreover, note that this definition is *independent* of P. In addition, since $t \mapsto \operatorname{dist}(t,A)$ is continuous for arbitrary $A \subset \mathbb{R}$ and $x \mapsto \operatorname{dist}(t,\mathcal{M}_{L,\mathrm{Q}_i,x}(0^+))$, $i \in \{1,2\}$, is measurable by Aumann's measurable selection principle stated in Lemma A.3.18, we see by Lemma A.3.17 that $\pi : \mathbb{R} \times X \to \mathbb{R}$ is measurable. Now a simple calculation shows

$$h(x)\mathcal{C}_{L_{\mathrm{class}},\bar{\mathrm{P}}(\,\cdot\,|x)}(\pi(t,x)) \;\leq\; \mathcal{C}_{L,\mathrm{P}(\,\cdot\,|x),x}(t) - \mathcal{C}^*_{L,\mathrm{P}(\,\cdot\,|x),x}\,, \qquad x \in X,\, t \in \mathbb{R},$$

and thus we find $\|h\|_{L_1(\mu)}\mathcal{R}_{L_{\mathrm{class}},\bar{\mathrm{P}}}(\pi \circ f) \leq \mathcal{R}_{L,\mathrm{P}}(f) - \mathcal{R}^*_{L,\mathrm{P}}$ for all measurable $f : X \to \mathbb{R}$, where $\pi \circ f(x) := \pi(f(x),x)$, $x \in X$. Now let $\mathcal{L}$ be a measurable learning method producing decision functions f_D. Then $\pi \circ \mathcal{L}$ defined by $D \mapsto \pi \circ f_D$ is also a measurable learning method since π is measurable. Moreover, $\pi \circ \mathcal{L}$ is independent of P. Assume that $\mathcal{L}$ learns all distributions P of type $\{\mathrm{Q}_1,\mathrm{Q}_2\}$ that satisfy $\mathrm{P}_X = \mu$ with a uniform rate. Then our considerations above show that $\pi \circ \mathcal{L}$ learns all associated classification problems $\bar{\mathrm{P}}$ with the same uniform rate, but by Corollary 6.7 this is impossible. □

The results above show that in general we cannot *a priori* guarantee with a fixed confidence that a learning method finds a nearly optimal decision function. This is a fundamental limitation for statistical learning methods that we *cannot* elude by, e.g., cleverly combining different learning methods since such a procedure itself constitutes a learning method. In other words, the only way to resolve this issue is to make *a priori* assumptions on the data generating distribution P. However, since in almost no case will we be able to *rigorously check* whether P actually satisfies the imposed assumptions, such an approach has only very limited utility for *a priori guaranteeing* good generalization performance. On the other hand, by establishing learning rates for different types of assumptions on P, we can *understand* for which kind of distributions the learning method considered learns easily and for which it does not. In turn, such knowledge can then be used in practice where one *has to* decide which learning methods are likely to be appropriate for a specific application.

6.2 Basic Concentration Inequalities

We will see in the following sections that our statistical analysis of both ERM and SVMs relies heavily on bounds on the probabilities

$$\mathrm{P}^n\big(\{D \in (X \times Y)^n : |\mathcal{R}_{L,\mathrm{D}}(f) - \mathcal{R}_{L,\mathrm{P}}(f)| > \varepsilon\}\big)\,.$$

In this section, we thus establish some basic bounds on such probabilities.

Let us begin with an elementary yet powerful inequality that will be the key ingredient for *all* the more advanced results that follow (see also Exercise 6.2 for a slightly refined estimate).

Theorem 6.9 (Markov's inequality). *Let $(\Omega, \mathcal{A}, \mathrm{P})$ be a probability space. Then, for all measurable functions $f : \Omega \to \mathbb{R}$ and all $t > 0$, we have*

$$\mathrm{P}\bigl(\{\omega \in \Omega : f(\omega) \geq t\}\bigr) \;\leq\; \frac{\mathbb{E}_{\mathrm{P}}|f|}{t}\,.$$

Proof. For $A_t := \{\omega \in \Omega : f(\omega) \geq t\}$, we obviously have $t\,\mathbf{1}_{A_t} \leq f\mathbf{1}_{A_t} \leq |f|$, and hence we obtain $t\,\mathrm{P}(A_t) = \mathbb{E}_{\mathrm{P}}\,t\,\mathbf{1}_{A_t} \leq \mathbb{E}_{\mathrm{P}}|f|$. □

From Markov's inequality, it is straightforward to derive Chebyshev's inequality $\mathrm{P}(\{\omega \in \Omega : |f(\omega)| \geq t\}) \leq t^{-2}\mathbb{E}_{\mathrm{P}}|f|^2$. The following result also follows from Markov's inequality.

Theorem 6.10 (Hoeffding's inequality). *Let $(\Omega, \mathcal{A}, \mathrm{P})$ be a probability space, $a < b$ be two real numbers, $n \geq 1$ be an integer, and $\xi_1, \dots, \xi_n : \Omega \to [a, b]$ be independent random variables. Then, for all $\tau > 0$, we have*

$$\mathrm{P}\left(\frac{1}{n}\sum_{i=1}^{n}\bigl(\xi_i - \mathbb{E}_{\mathrm{P}}\xi_i\bigr) \geq (b-a)\sqrt{\frac{\tau}{2n}}\right) \leq e^{-\tau}\,.$$

Proof. We begin with a preliminary consideration. To this end, let $\tilde{a} < \tilde{b}$ be two real numbers and $\xi : \Omega \to [\tilde{a}, \tilde{b}]$ be a random variable with $\mathbb{E}_{\mathrm{P}}\xi = 0$. Note that from this assumptions we can immediately conclude that $\tilde{a} \leq 0$ and $\tilde{b} \geq 0$. Moreover, observe that for $x \in [\tilde{a}, \tilde{b}]$ we have

$$x = \frac{\tilde{b} - x}{\tilde{b} - \tilde{a}}\tilde{a} + \frac{x - \tilde{a}}{\tilde{b} - \tilde{a}}\tilde{b}\,,$$

and hence the convexity of the exponential function implies

$$e^{tx} \leq \frac{\tilde{b} - x}{\tilde{b} - \tilde{a}}e^{t\tilde{a}} + \frac{x - \tilde{a}}{\tilde{b} - \tilde{a}}e^{t\tilde{b}}\,, \qquad t > 0.$$

Since $\mathbb{E}_{\mathrm{P}}\xi = 0$, we then obtain

$$\begin{aligned}\mathbb{E}_{\mathrm{P}}e^{t\xi} \leq \mathbb{E}_{\mathrm{P}}\left(\frac{\tilde{b} - \xi}{\tilde{b} - \tilde{a}}e^{t\tilde{a}} + \frac{\xi - \tilde{a}}{\tilde{b} - \tilde{a}}e^{t\tilde{b}}\right) &= \frac{\tilde{b}}{\tilde{b} - \tilde{a}}e^{t\tilde{a}} - \frac{\tilde{a}}{\tilde{b} - \tilde{a}}e^{t\tilde{b}} \\ &= e^{t\tilde{a}}\left(1 + \frac{\tilde{a}}{\tilde{b} - \tilde{a}} - \frac{\tilde{a}}{\tilde{b} - \tilde{a}}e^{t(\tilde{b} - \tilde{a})}\right)\end{aligned}$$

for all $t > 0$. Let us now write $p := -\tilde{a}\,(\tilde{b} - \tilde{a})^{-1}$. Then we observe that $\tilde{a} \leq 0$ implies $p \geq 0$, and $\tilde{b} \geq 0$ implies $p \leq 1$. For $s \in \mathbb{R}$, we hence find $e^s > 0 \geq 1 - 1/p$, from which we conclude that $1 - p + pe^s > 0$. Consequently, $\phi_p(s) := \ln(1 - p + pe^s) - ps$ is defined for all $s \in \mathbb{R}$. Moreover, these definitions together with our previous estimate yield

$$\mathbb{E}_{\mathrm{P}}e^{t\xi} \leq e^{-tp(\tilde{b} - \tilde{a})}\left(1 - p + pe^{t(\tilde{b} - \tilde{a})}\right) = e^{\phi_p(t(\tilde{b} - \tilde{a}))}\,.$$

Now observe that we have $\phi_p(0) = 0$ and $\phi_p'(s) = \frac{pe^s}{1-p+pe^s} - p$. From the latter, we conclude that $\phi_p'(0) = 0$ and

$$\phi_p''(s) = \frac{(1-p+pe^s)pe^s - pe^s pe^s}{(1-p+pe^s)^2} = \frac{(1-p)pe^s}{(1-p+pe^s)^2} \leq \frac{(1-p)pe^s}{4(1-p)pe^s} = \frac{1}{4}$$

for all $s \in \mathbb{R}$. By Taylor's formula with Lagrangian remainder, we hence find

$$\phi_p(s) = \phi_p(0) + \phi_p'(0)s + \frac{1}{2}\phi_p''(s')s^2 \leq \frac{s^2}{8}\,, \qquad s > 0,$$

where $s' \in [0, s]$ is a suitable real number. Consequently, we obtain

$$\mathbb{E}_\mathrm{P} e^{t\xi} \leq e^{\phi_p(t(\tilde{b}-\tilde{a}))} \leq \exp\biggl(\frac{t^2(\tilde{b}-\tilde{a})^2}{8}\biggr)\,, \qquad t > 0\,.$$

Applying this estimate to the random variables $\xi_i - \mathbb{E}\xi_i : \Omega \to [a - \mathbb{E}\xi_i, b - \mathbb{E}\xi_i]$, where $\mathbb{E} := \mathbb{E}_\mathrm{P}$, we now find

$$\mathbb{E}_\mathrm{P} e^{t(\xi_i - \mathbb{E}\xi_i)} \leq \exp\biggl(\frac{t^2(b-a)^2}{8}\biggr)\,, \qquad t > 0,\ i = 1, \dots, n.$$

Using this estimate together with Markov's inequality and the independence assumption, we hence obtain with $\mathbb{E} := \mathbb{E}_\mathrm{P}$ that

$$\begin{aligned}\mathrm{P}\biggl(\sum_{i=1}^n (\xi_i - \mathbb{E}\xi_i) \geq \varepsilon n\biggr) \leq e^{-t\varepsilon n}\,\mathbb{E}\exp\biggl(t\sum_{i=1}^n (\xi_i - \mathbb{E}\xi_i)\biggr) &\leq e^{-t\varepsilon n}\prod_{i=1}^n \mathbb{E}e^{t(\xi_i - \mathbb{E}\xi_i)} \\ &\leq e^{-t\varepsilon n}\, e^{\frac{nt^2(b-a)^2}{8}}\end{aligned}$$

for all $\varepsilon > 0$ and $t > 0$. Now we obtain the assertion by considering $\varepsilon := (b-a)(\frac{\tau}{2n})^{1/2}$ and $t := \frac{4\varepsilon}{(b-a)^2}$. □

Our next goal is to present a concentration inequality that refines Hoeffding's inequality when we know not only the $\|\cdot\|_\infty$-norms of the random variables involved but also their variances, i.e., their $\|\cdot\|_2$-norms. To this end, we need the following technical lemma.

Lemma 6.11. *For all $x > -1$, we have* $(1+x)\ln(1+x) - x \geq \frac{3}{2}\frac{x^2}{x+3}$.

Proof. For $x > -1$, we define $f(x) := (1+x)\ln(1+x) - x$ and $g(x) := \frac{3}{2}\frac{x^2}{x+3}$. Then an easy calculation shows that, for all $x > -1$, we have

$$f'(x) = \ln(1+x)\,, \qquad f''(x) = \frac{1}{1+x}\,,$$

$$g'(x) = \frac{3x^2 + 18x}{2(x+3)^2}\,, \qquad g''(x) = \frac{27}{(x+3)^3}\,.$$

Consequently, we have $f(0) = g(0) = 0$, $f'(0) = g'(0) = 0$, and $f''(x) \geq g''(x)$ for all $x > -1$. For $x \geq 0$, the fundamental theorem of calculus thus gives

$$f'(x) = \int_0^x f''(t)dt \geq \int_0^x g''(t)dt = g'(x)\,,$$

and by repeating this reasoning, we obtain the assertion for $x \geq 0$. For $x \in (-1, 0]$, we can show the assertion analogously. □

Now we can establish the announced refinement of Hoeffding's inequality.

Theorem 6.12 (Bernstein's inequality). *Let* $(\Omega, \mathcal{A}, \mathrm{P})$ *be a probability space,* $B > 0$ *and* $\sigma > 0$ *be real numbers, and* $n \geq 1$ *be an integer. Furthermore, let* $\xi_1, \ldots, \xi_n : \Omega \to \mathbb{R}$ *be independent random variables satisfying* $\mathbb{E}_\mathrm{P}\xi_i = 0$*,* $\|\xi_i\|_\infty \leq B$*, and* $\mathbb{E}_\mathrm{P}\xi_i^2 \leq \sigma^2$ *for all* $i = 1, \ldots, n$*. Then we have*

$$\mathrm{P}\biggl(\frac{1}{n}\sum_{i=1}^n \xi_i \geq \sqrt{\frac{2\sigma^2\tau}{n}} + \frac{2B\tau}{3n}\biggr) \leq e^{-\tau}\,, \qquad\qquad \tau > 0.$$

Proof. By Markov's inequality and the independence of $\xi_1, \ldots, \xi_n$, we have

$$\mathrm{P}\biggl(\sum_{i=1}^n \xi_i \geq \varepsilon n\biggr) \leq e^{-t\varepsilon n}\,\mathbb{E}_\mathrm{P} \exp\biggl(t\sum_{i=1}^n \xi_i\biggr) \leq e^{-t\varepsilon n}\prod_{i=1}^n \mathbb{E}_\mathrm{P}\, e^{t\xi_i}$$

for all $t \geq 0$ and $\varepsilon > 0$. Furthermore, the properties of ξ_i imply

$$\mathbb{E}_\mathrm{P}\, e^{t\xi_i} = \sum_{k=0}^\infty \frac{t^k}{k!}\mathbb{E}_\mathrm{P}\xi_i^k \leq 1 + \sum_{k=2}^\infty \frac{t^k}{k!}\sigma^2 B^{k-2} = 1 + \frac{\sigma^2}{B^2}\bigl(e^{tB} - tB - 1\bigr)\,.$$

Using the simple estimate $1 + x \leq e^x$ for $x := \frac{\sigma^2}{B^2}(e^{tB} - tB - 1)$, we hence obtain

$$\begin{aligned}\mathrm{P}\biggl(\sum_{i=1}^n \xi_i \geq \varepsilon n\biggr) &\leq e^{-t\varepsilon n}\prod_{i=1}^n\Bigl(1 + \frac{\sigma^2}{B^2}\bigl(e^{tB} - tB - 1\bigr)\Bigr) \\ &\leq \exp\Bigl(-t\varepsilon n + \frac{\sigma^2 n}{B^2}\bigl(e^{tB} - tB - 1\bigr)\Bigr)\end{aligned}$$

for all $t \geq 0$. Now elementary calculus shows that the right-hand side of the inequality is minimized at

$$t^* := \frac{1}{B}\ln\Bigl(1 + \frac{\varepsilon B}{\sigma^2}\Bigr)\,.$$

Writing $y := \frac{\varepsilon B}{\sigma^2}$ and using Lemma 6.11, we furthermore obtain

$$\begin{aligned}-t^*\varepsilon n + \frac{\sigma^2 n}{B^2}\bigl(e^{t^*B} - t^*B - 1\bigr) = -\frac{n\sigma^2}{B^2}\bigl((1+y)\ln(1+y) - y\bigr) &\leq -\frac{3n\sigma^2}{2B^2}\frac{y^2}{y+3} \\ &= -\frac{3\varepsilon^2 n}{2\varepsilon B + 6\sigma^2}\,.\end{aligned}$$

Let us now define $\varepsilon := \sqrt{\frac{2\sigma^2\tau}{n} + \frac{B^2\tau^2}{9n^2}} + \frac{B\tau}{3n}$. Then we have $\tau = \frac{3\varepsilon^2 n}{2\varepsilon B + 6\sigma^2}$ and

$$\varepsilon \leq \sqrt{\frac{2\sigma^2\tau}{n}} + \frac{2B\tau}{3n}\,,$$

and thus we obtain the assertion. □

It is important to keep in mind that in situations where we know upper bounds σ^2 and B for the variances and the suprema, respectively, Bernstein's inequality is often sharper than Hoeffding's inequality. The details are discussed in Exercise 6.1.

Our next goal is to generalize Bernstein's inequality to Hilbert space valued random variables. To this end, we need the following more general result.

Theorem 6.13. *Let $(\Omega, \mathcal{A}, \mathrm{P})$ be a probability space, E be a separable Banach space, and $\xi_1, \ldots, \xi_n : \Omega \to E$ be independent E-valued* P*-integrable random variables. Then, for all $\varepsilon > 0$ and all $t \geq 0$, we have*

$$\mathrm{P}\biggl(\Bigl\|\sum_{i=1}^n \xi_i\Bigr\| \geq \varepsilon n\biggr) \leq \exp\biggl(-t\varepsilon n + t\mathbb{E}_\mathrm{P}\Bigl\|\sum_{i=1}^n \xi_i\Bigr\| + \sum_{i=1}^n \mathbb{E}_\mathrm{P}\bigl(e^{t\|\xi_i\|} - 1 - t\|\xi_i\|\bigr)\biggr)\,.$$

Proof. Let us consider the σ-algebras $\mathcal{F}_0 := \{\emptyset, \Omega\}$ and $\mathcal{F}_k := \sigma(\xi_1, \ldots, \xi_k)$, $k = 1, \ldots, n$. Furthermore, for $k = 1, \ldots, n$, we define

$$\begin{aligned} X_k &:= \mathbb{E}_\mathrm{P}\biggl(\Bigl\|\sum_{i=1}^n \xi_i\Bigr\| \,\Big|\, \mathcal{F}_k\biggr) - \mathbb{E}_\mathrm{P}\biggl(\Bigl\|\sum_{i=1}^n \xi_i\Bigr\| \,\Big|\, \mathcal{F}_{k-1}\biggr)\,, \\ Y_k &:= \mathbb{E}_\mathrm{P}\biggl(\Bigl\|\sum_{i=1}^n \xi_i\Bigr\| - \Bigl\|\sum_{i\neq k} \xi_i\Bigr\| \,\Big|\, \mathcal{F}_k\biggr)\,. \end{aligned}$$

By a simple telescope sum argument, we then have

$$\sum_{i=1}^n X_i = \Bigl\|\sum_{i=1}^n \xi_i\Bigr\| - \mathbb{E}_\mathrm{P}\Bigl\|\sum_{i=1}^n \xi_i\Bigr\|\,. \tag{6.6}$$

Moreover, note that $\sum_{i\neq k} \xi_i$ is independent of ξ_k, and hence we obtain

$$\mathbb{E}_\mathrm{P}\biggl(\Bigl\|\sum_{i\neq k} \xi_i\Bigr\| \,\Big|\, \mathcal{F}_k\biggr) = \mathbb{E}_\mathrm{P}\biggl(\Bigl\|\sum_{i\neq k} \xi_i\Bigr\| \,\Big|\, \mathcal{F}_{k-1}, \xi_k\biggr) = \mathbb{E}_\mathrm{P}\biggl(\Bigl\|\sum_{i\neq k} \xi_i\Bigr\| \,\Big|\, \mathcal{F}_{k-1}\biggr)\,.$$

Since $\mathcal{F}_{k-1} \subset \mathcal{F}_k$, we thus find

$$\begin{aligned} X_k &= \mathbb{E}_\mathrm{P}\biggl(\Bigl\|\sum_{i=1}^n \xi_i\Bigr\| \,\Big|\, \mathcal{F}_k\biggr) - \mathbb{E}_\mathrm{P}\biggl(\Bigl\|\sum_{i=1}^n \xi_i\Bigr\| \,\Big|\, \mathcal{F}_{k-1}\biggr) \\ &= \mathbb{E}_\mathrm{P}\biggl(\Bigl\|\sum_{i=1}^n \xi_i\Bigr\| - \Bigl\|\sum_{i\neq k} \xi_i\Bigr\| \,\Big|\, \mathcal{F}_k\biggr) - \mathbb{E}_\mathrm{P}\biggl(\Bigl\|\sum_{i=1}^n \xi_i\Bigr\| - \Bigl\|\sum_{i\neq k} \xi_i\Bigr\| \,\Big|\, \mathcal{F}_{k-1}\biggr) \\ &= Y_k - \mathbb{E}_\mathrm{P}(Y_k|\mathcal{F}_{k-1}) \end{aligned} \tag{6.7}$$

for all $k = 1, \dots, n$. Using $x \le e^{x-1}$ for all $x \in \mathbb{R}$, we hence obtain

$$\begin{aligned}\mathbb{E}_{\mathrm{P}}\bigl(e^{tX_k}\,|\,\mathcal{F}_{k-1}\bigr) &= e^{-t\mathbb{E}_{\mathrm{P}}(Y_k|\mathcal{F}_{k-1})}\mathbb{E}_{\mathrm{P}}\bigl(e^{tY_k}\,|\,\mathcal{F}_{k-1}\bigr)\\ &\le e^{-t\mathbb{E}_{\mathrm{P}}(Y_k|\mathcal{F}_{k-1})}e^{\mathbb{E}_{\mathrm{P}}(e^{tY_k}\,|\,\mathcal{F}_{k-1})-1}\\ &= \exp\Bigl(\mathbb{E}_{\mathrm{P}}\bigl(e^{tY_k}-1-tY_k\,|\,\mathcal{F}_{k-1}\bigr)\Bigr)\,. \qquad (6.8)\end{aligned}$$

Now, an easy calculation shows $e^x - e^{-x} \ge 2x$ for all $x \ge 0$, which in turn implies $e^{-x} - 1 - (-x) \le e^x - 1 - x$ for all $x \ge 0$. From this we conclude that $e^x - 1 - x \le e^{|x|} - 1 - |x|$ for all $x \in \mathbb{R}$. Moreover, it is straightforward to check that the function $x \mapsto e^x - 1 - x$ is increasing on $[0, \infty)$. In addition, the triangle inequality in E gives

$$|Y_k| \le \mathbb{E}_{\mathrm{P}}\Biggl(\biggl|\Bigl\|\sum_{i=1}^n \xi_i\Bigr\| - \Bigl\|\sum_{i\neq k}\xi_i\Bigr\|\biggr|\,\Bigg|\,\mathcal{F}_k\Biggr) \le \mathbb{E}_{\mathrm{P}}\Biggl(\biggl\|\sum_{i=1}^n \xi_i - \sum_{i\neq k}\xi_i\biggr\|\,\Bigg|\,\mathcal{F}_k\Biggr) = \|\xi_k\|\,,$$

where in the last step we used that $\|\xi_k\|$ is $\mathcal{F}_k$-measurable. Consequently, (6.8) implies

$$\begin{aligned}\mathbb{E}_{\mathrm{P}}\bigl(e^{tX_k}\,|\,\mathcal{F}_{k-1}\bigr) &\le \exp\Bigl(\mathbb{E}_{\mathrm{P}}\bigl(e^{t|Y_k|}-1-t|Y_k|\,\big|\,\mathcal{F}_{k-1}\bigr)\Bigr)\\ &\le \exp\Bigl(\mathbb{E}_{\mathrm{P}}\bigl(e^{t\|\xi_k\|}-1-t\|\xi_k\|\,\big|\,\mathcal{F}_{k-1}\bigr)\Bigr)\\ &= \exp\Bigl(\mathbb{E}_{\mathrm{P}}\bigl(e^{t\|\xi_k\|}-1-t\|\xi_k\|\bigr)\Bigr)\,, \qquad (6.9)\end{aligned}$$

where in the last step we used that ξ_k is independent of $\mathcal{F}_{k-1}$. Moreover, $\sum_{i=1}^{k-1} X_i$ is $\mathcal{F}_{k-1}$-measurable, and writing $\mathbb{E} := \mathbb{E}_{\mathrm{P}}$ we hence have

$$\mathbb{E}\Bigl(e^{t\sum_{i=1}^{k-1}X_i}\mathbb{E}\bigl(e^{tX_k}\,\big|\mathcal{F}_{k-1}\bigr)\Bigr) = \mathbb{E}\Bigl(\mathbb{E}\bigl(e^{t\sum_{i=1}^{k-1}X_i}e^{tX_k}\,\big|\mathcal{F}_{k-1}\bigr)\Bigr) = \mathbb{E}e^{t\sum_{i=1}^{k}X_i}\,.$$

Combining this last equation with (6.9) now yields

$$\begin{aligned}\mathbb{E}_{\mathrm{P}}e^{t\sum_{i=1}^{k}X_i} &= \mathbb{E}_{\mathrm{P}}\Bigl(e^{t\sum_{i=1}^{k-1}X_i}\,\mathbb{E}_{\mathrm{P}}\bigl(e^{tX_k}\,\big|\mathcal{F}_{k-1}\bigr)\Bigr)\\ &\le \mathbb{E}_{\mathrm{P}}\Bigl(e^{t\sum_{i=1}^{k-1}X_i}\Bigr)\cdot\exp\Bigl(\mathbb{E}_{\mathrm{P}}\bigl(e^{t\|\xi_k\|}-1-t\|\xi_k\|\bigr)\Bigr)\,,\end{aligned}$$

and by successively applying this inequality we hence obtain

$$\mathbb{E}_{\mathrm{P}}e^{t\sum_{i=1}^{n}X_i} \le \prod_{i=1}^n \exp\Bigl(\mathbb{E}_{\mathrm{P}}\bigl(e^{t\|\xi_i\|}-1-t\|\xi_k\|\bigr)\Bigr) = \exp\Biggl(\sum_{i=1}^n \mathbb{E}_{\mathrm{P}}\bigl(e^{t\|\xi_i\|}-1-t\|\xi_i\|\bigr)\Biggr)$$

for all $t \ge 0$. By Markov's inequality and (6.6), we thus find

$$\begin{aligned}
\mathrm{P}\biggl(\Bigl\|\sum_{i=1}^n \xi_i\Bigr\| \geq \varepsilon n\biggr) &\leq e^{-t\varepsilon n}\,\mathbb{E}_\mathrm{P}\exp\biggl(t\Bigl\|\sum_{i=1}^n \xi_i\Bigr\|\biggr)\\
&= e^{-t\varepsilon n}\,\mathbb{E}_\mathrm{P}\exp\biggl(t\sum_{i=1}^n X_i + t\mathbb{E}_\mathrm{P}\Bigl\|\sum_{i=1}^n \xi_i\Bigr\|\biggr)\\
&\leq \exp\biggl(-t\varepsilon n + t\mathbb{E}_\mathrm{P}\Bigl\|\sum_{i=1}^n \xi_i\Bigr\| + \sum_{i=1}^n \mathbb{E}_\mathrm{P}\bigl(e^{t\|\xi_i\|} - 1 - t\|\xi_i\|\bigr)\biggr)
\end{aligned}$$

for all $\varepsilon > 0$ and all $t \geq 0$. □

With the help of the previous theorem we can now establish the following Hilbert space version of Bernstein's inequality.

Theorem 6.14 (Bernstein's inequality in Hilbert spaces). *Let $(\Omega, \mathcal{A}, \mathrm{P})$ be a probability space, H be a separable Hilbert space, $B > 0$, and $\sigma > 0$. Furthermore, let $\xi_1, \dots, \xi_n : \Omega \to H$ be independent random variables satisfying $\mathbb{E}_\mathrm{P}\xi_i = 0$, $\|\xi_i\|_\infty \leq B$, and $\mathbb{E}_\mathrm{P}\|\xi_i\|_H^2 \leq \sigma^2$ for all $i = 1, \dots, n$. Then we have*

$$\mathrm{P}\biggl(\Bigl\|\frac{1}{n}\sum_{i=1}^n \xi_i\Bigr\|_H \geq \sqrt{\frac{2\sigma^2\tau}{n}} + \sqrt{\frac{\sigma^2}{n}} + \frac{2B\tau}{3n}\biggr) \;\leq\; e^{-\tau}\,, \qquad \tau > 0.$$

Proof. We will prove the assertion by applying Theorem 6.13. To this end, we first observe that the independence of $\xi_1, \dots, \xi_n$ yields $\mathbb{E}\langle\xi_i, \xi_j\rangle_H = \langle\mathbb{E}\xi_i, \mathbb{E}\xi_j\rangle_H = 0$ for all $i \neq j$, where $\mathbb{E} := \mathbb{E}_\mathrm{P}$. Consequently, we obtain

$$\mathbb{E}\Bigl\|\sum_{i=1}^n \xi_i\Bigr\|_H \leq \biggl(\mathbb{E}\Bigl\|\sum_{i=1}^n \xi_i\Bigr\|_H^2\biggr)^{1/2} = \biggl(\sum_{i=1}^n \mathbb{E}\|\xi_i\|_H^2\biggr)^{1/2} \leq \sqrt{n\sigma^2}\,. \tag{6.10}$$

In addition, the series expansion of the exponential function yields

$$\sum_{i=1}^n \mathbb{E}\bigl(e^{t\|\xi_i\|} - 1 - t\|\xi_i\|\bigr) = \sum_{i=1}^n\sum_{j=2}^\infty \frac{t^j}{j!}\mathbb{E}\|\xi_i\|_H^j \leq \sum_{i=1}^n\sum_{j=2}^\infty \frac{t^j}{j!}B^{j-2}\mathbb{E}\|\xi_i\|_H^2$$

for all $t \geq 0$, and therefore we find

$$\sum_{i=1}^n \mathbb{E}\bigl(e^{t\|\xi_i\|} - 1 - t\|\xi_i\|\bigr) \leq \frac{\sigma^2}{B^2}\sum_{i=1}^n\sum_{j=2}^\infty \frac{t^j}{j!}B^j = \frac{n\sigma^2}{B^2}\bigl(e^{tB} - 1 - tB\bigr)$$

for all $t \geq 0$. By Theorem 6.13, we hence obtain

$$\begin{aligned}
\mathrm{P}\biggl(\Bigl\|\sum_{i=1}^n \xi_i\Bigr\|_H \geq \varepsilon n\biggr) &\leq \exp\biggl(-t\varepsilon n + t\mathbb{E}\Bigl\|\sum_{i=1}^n \xi_i\Bigr\| + \sum_{i=1}^n \mathbb{E}(e^{t\|\xi_i\|} - 1 - t\|\xi_i\|)\biggr)\\
&\leq \exp\Bigl(-t\varepsilon n + t\sqrt{n\sigma^2} + \frac{n\sigma^2}{B^2}\bigl(e^{tB} - 1 - tB\bigr)\Bigr) \qquad (6.11)
\end{aligned}$$

for all $t \geq 0$. Let us now restrict our considerations to $\varepsilon \geq \sigma n^{-1/2}$. Then we have $y := \frac{\varepsilon B}{\sigma^2} - \frac{B}{\sqrt{n\sigma^2}} > 0$, and consequently it is easy to see that the function on the right-hand side of (6.11) is minimized at

$$t^* := \frac{1}{B} \ln\Bigl(1 + \frac{\varepsilon B}{\sigma^2} - \frac{B}{\sqrt{n\sigma^2}}\Bigr) .$$

Moreover, Lemma 6.11 yields

$$\begin{aligned} -t^*\varepsilon n + t^*\sqrt{n\sigma^2} + \frac{n\sigma^2}{B^2}\bigl(e^{t^*B} - 1 - t^*B\bigr) &= -\frac{n\sigma^2}{B^2}\Bigl((1+y)\ln(1+y) - y\Bigr) \\ &\leq -\frac{3n\sigma^2}{2B^2}\frac{y^2}{y+3} . \end{aligned}$$

By combining this estimate with (6.11), we then find

$$\mathrm{P}\Bigl(\Bigl\| \sum_{i=1}^n \xi_i \Bigr\|_H \geq \varepsilon n\Bigr) \;\leq\; \exp\Bigl(-\frac{3n\sigma^2}{2B^2}\frac{y^2}{y+3}\Bigr) .$$

Let us now define $\varepsilon := \sqrt{\frac{2\sigma^2\tau}{n} + \frac{B^2\tau^2}{9n^2}} + \frac{B\tau}{3n} + \sqrt{\frac{\sigma^2}{n}}$. Then, an easy calculation shows

$$y = \frac{\varepsilon B}{\sigma^2} - \frac{B}{\sqrt{n\sigma^2}} = \sqrt{\frac{2B^2\tau}{n\sigma^2} + \frac{B^4\tau^2}{9n^2\sigma^4}} + \frac{B^2\tau}{3n\sigma^2} ,$$

and hence we find $\tau = -\frac{3n\sigma^2}{2B^2}\frac{y^2}{y+3}$. Now the assertion follows from

$$\varepsilon = \sqrt{\frac{2\sigma^2\tau}{n} + \frac{B^2\tau^2}{9n^2}} + \frac{B\tau}{3n} + \sqrt{\frac{\sigma^2}{n}} \leq \sqrt{\frac{2\sigma^2\tau}{n}} + \frac{2B\tau}{3n} + \sqrt{\frac{\sigma^2}{n}} . \qquad \square$$

The following Hilbert space version of Hoeffding's inequality is an immediate consequence of Theorem 6.14. We will use it in Section 6.4 to derive an oracle inequality for SVMs.

Corollary 6.15 (Hoeffding's inequality in Hilbert spaces). *Let $(\Omega, \mathcal{A}, \mathrm{P})$ be a probability space, H be a separable Hilbert space H, and $B > 0$. Furthermore, let $\xi_1, \ldots, \xi_n : \Omega \to H$ be independent H-valued random variables satisfying $\|\xi_i\|_\infty \leq B$ for all $i = 1, \ldots, n$. Then, for all $\tau > 0$, we have*

$$\mathrm{P}\Bigl(\Bigl\| \frac{1}{n}\sum_{i=1}^n \bigl(\xi_i - \mathbb{E}_{\mathrm{P}}\xi_i\bigr) \Bigr\|_H \geq B\sqrt{\frac{2\tau}{n}} + B\sqrt{\frac{1}{n}} + \frac{4B\tau}{3n}\Bigr) \leq e^{-\tau} .$$

Proof. Let us define $\eta_i := \xi_i - \mathbb{E}_{\mathrm{P}}\xi_i$, $i = 1, \ldots, n$. Then we have $\mathbb{E}_{\mathrm{P}}\eta_i = 0$, $\|\eta_i\|_\infty \leq 2B$, and

$$\mathbb{E}_{\mathrm{P}}\|\eta_i\|_H^2 = \mathbb{E}_{\mathrm{P}}\langle \xi_i, \xi_i\rangle - 2\mathbb{E}_{\mathrm{P}}\langle \xi_i, \mathbb{E}_{\mathrm{P}}\xi_i\rangle + \langle \mathbb{E}_{\mathrm{P}}\xi_i, \mathbb{E}_{\mathrm{P}}\xi_i\rangle \leq \mathbb{E}_{\mathrm{P}}\langle \xi_i, \xi_i\rangle \leq B^2$$

for all $i = 1, \ldots, n$. Applying Theorem 6.14 to $\eta_1, \ldots, \eta_n$ then yields the assertion. $\square$

6.3 Statistical Analysis of Empirical Risk Minimization

In this section, we investigate statistical properties for empirical risk minimization. Although this learning method is not our primary object of interest, there are two reasons why we consider it before investigating SVMs. First, the method is so elementary that the basic ideas of its analysis are not hidden by technical considerations. This will give us good preparation for the more involved analysis of SVMs in Section 6.4. Second, the results we establish will be utilized in Section 6.5, where we investigate how the regularization parameter of SVMs can be chosen in a data-dependent and adaptive way.

Let us begin by formally introducing empirical risk minimization.

Definition 6.16. *Let $L : X \times Y \times \mathbb{R} \to [0, \infty)$ be a loss and $\mathcal{F} \subset \mathcal{L}_0(X)$ be a non-empty set. A learning method whose decision functions f_D satisfy*

$$\mathcal{R}_{L,\mathrm{D}}(f_D) = \inf_{f \in \mathcal{F}} \mathcal{R}_{L,\mathrm{D}}(f) \tag{6.12}$$

*for all $n \geq 1$ and $D \in (X \times Y)^n$ is called **empirical risk minimization (ERM)** with respect to L and $\mathcal{F}$.*

By definition, empirical risk minimization produces decision functions that minimize the empirical risk over $\mathcal{F}$. The *motivation* for this approach is based on the law of large numbers which says that for *fixed* $f \in \mathcal{F}$ we have

$$\lim_{n \to \infty} \mathcal{R}_{L,\mathrm{D}}(f) = \mathcal{R}_{L,\mathrm{P}}(f)$$

if the training sets D of length n are identically and independently distributed according to some probability measure P on $X \times Y$. This limit relation suggests that in order to find a minimizer of the true risk $\mathcal{R}_{L,\mathrm{P}}$, it suffices to find a minimizer of its empirical approximation $\mathcal{R}_{L,\mathrm{D}}$. Unfortunately, however, minimizing $\mathcal{R}_{L,\mathrm{D}}$ over $\mathcal{F} := \mathcal{L}_0(X)$ or $\mathcal{F} := \mathcal{L}_\infty(X)$ can lead to "overfitted" decision functions, as discussed in Exercise 6.7, and hence ERM typically minimizes over a smaller set $\mathcal{F}$ of functions. Moreover, note that for general losses L and sets of functions $\mathcal{F}$ there does not necessarily exist a function f_D satisfying (6.12). In addition, there are also situations in which multiple minimizers exist, and consequently one should always be aware that ERM usually is not a uniquely determined learning method. Let us now show that there usually exists a measurable ERM if there exists an ERM.[1]

Lemma 6.17 (Measurability of ERM). *Let $L : X \times Y \times \mathbb{R} \to [0, \infty)$ be a loss and $\mathcal{F} \subset \mathcal{L}_0(X)$ be a subset that is equipped with a complete and separable metric dominating the pointwise convergence. Then, if there exists an ERM, there also exists a measurable ERM.*

[1] Since this is little more than a technical requirement for the following results, the first-time reader may skip this lemma.

Before we present the proof of this lemma, let us first note that *finite* subsets $\mathcal{F} \subset \mathcal{L}_0(X)$ equipped with the discrete metric satisfy the assumptions above. In addition, the existence of an ERM is automatically guaranteed in this case, and hence there always exists a measurable ERM for finite $\mathcal{F}$. Similarly, for *closed, separable* $\mathcal{F} \subset \mathcal{L}_\infty(X)$, there exists a measurable ERM whenever there exists an ERM.

Proof. Lemma 2.11 shows that the map $(x, y, f) \mapsto L(x, y, f(x))$ defined on $X \times Y \times \mathcal{F}$ is measurable. From this it is easy to conclude that the map $\varphi : (X \times Y)^n \times \mathcal{F} \to [0, \infty)$ defined by

$$\varphi(D, f) := \mathcal{R}_{L,\mathrm{D}}(f)\,, \qquad\qquad D \in (X \times Y)^n,\ f \in \mathcal{F},$$

is measurable with respect to the product topology of $(X \times Y)^n \times \mathcal{F}$. By taking $F(D) := \mathcal{F}$, $D \in (X \times Y)^n$, in Aumann's measurable selection principle (see Lemma A.3.18), we thus see that there exists an ERM such that $D \mapsto f_D$ is measurable with respect to the universal completion of the product σ-algebra of $(X \times Y)^n$. Consequently, the map $(X \times Y)^n \times X \to \mathcal{F} \times X$ defined by $(D, x) \mapsto (f_D, x)$ is measurable with respect to the universal completion of the product σ-algebra of $(X \times Y)^n \times X$. In addition, Lemma 2.11 shows that the map $\mathcal{F} \times X \to \mathbb{R}$ defined by $(f, x) \mapsto f(x)$ is measurable. Combining both maps, we then obtain the measurability of $(D, x) \mapsto f_D(x)$. □

Let us now analyze the statistical properties of ERM. To this end, let us assume that $\mathcal{R}^*_{L,\mathrm{P},\mathcal{F}} < \infty$. Moreover, let us fix a $\delta > 0$ and a function $f_\delta \in \mathcal{F}$ such that $\mathcal{R}_{L,\mathrm{P}}(f_\delta) \leq \mathcal{R}^*_{L,\mathrm{P},\mathcal{F}} + \delta$. Then a simple calculation shows

$$\begin{aligned}\mathcal{R}_{L,\mathrm{P}}(f_D) - \mathcal{R}^*_{L,\mathrm{P},\mathcal{F}} &\leq \mathcal{R}_{L,\mathrm{P}}(f_D) - \mathcal{R}_{L,\mathrm{D}}(f_D) + \mathcal{R}_{L,\mathrm{D}}(f_D) - \mathcal{R}_{L,\mathrm{P}}(f_\delta) + \delta \\ &\leq \mathcal{R}_{L,\mathrm{P}}(f_D) - \mathcal{R}_{L,\mathrm{D}}(f_D) + \mathcal{R}_{L,\mathrm{D}}(f_\delta) - \mathcal{R}_{L,\mathrm{P}}(f_\delta) + \delta \\ &\leq 2 \sup_{f \in \mathcal{F}} \bigl|\mathcal{R}_{L,\mathrm{P}}(f) - \mathcal{R}_{L,\mathrm{D}}(f)\bigr| + \delta\,,\end{aligned}$$

and by letting $\delta \to 0$ we thus find

$$\mathcal{R}_{L,\mathrm{P}}(f_D) - \mathcal{R}^*_{L,\mathrm{P},\mathcal{F}} \ \leq\ 2 \sup_{f \in \mathcal{F}} \bigl|\mathcal{R}_{L,\mathrm{P}}(f) - \mathcal{R}_{L,\mathrm{D}}(f)\bigr|\,. \tag{6.13}$$

Let us now assume that $\mathcal{F}$ is a finite set with cardinality $|\mathcal{F}|$ and that $B > 0$ is a real number such that

$$L(x, y, f(x)) \leq B\,, \qquad\qquad (x, y) \in X \times Y,\ f \in \mathcal{F}. \tag{6.14}$$

Note that the latter assumption ensures the earlier imposed $\mathcal{R}^*_{L,\mathrm{P},\mathcal{F}} < \infty$. For a measurable ERM, (6.13) together with Hoeffding's inequality then yields

$$\begin{aligned}
&\mathrm{P}^n\biggl(D \in (X\times Y)^n : \mathcal{R}_{L,\mathrm{P}}(f_D) - \mathcal{R}^*_{L,\mathrm{P},\mathcal{F}} \geq B\sqrt{\frac{2\tau}{n}}\,\biggr)\\
&\leq \mathrm{P}^n\biggl(D \in (X\times Y)^n : \sup_{f\in\mathcal{F}}\bigl|\mathcal{R}_{L,\mathrm{P}}(f) - \mathcal{R}_{L,\mathrm{D}}(f)\bigr| \geq B\sqrt{\frac{\tau}{2n}}\,\biggr)\\
&\leq \sum_{f\in\mathcal{F}} \mathrm{P}^n\biggl(D \in (X\times Y)^n : \bigl|\mathcal{R}_{L,\mathrm{P}}(f) - \mathcal{R}_{L,\mathrm{D}}(f)\bigr| \geq B\sqrt{\frac{\tau}{2n}}\,\biggr)\\
&\leq 2\,|\mathcal{F}|e^{-\tau}\,. \qquad (6.15)
\end{aligned}$$

By elementary algebraic transformations, we thus find the following result.

Proposition 6.18 (Oracle inequality for ERM). *Let $L : X \times Y \times \mathbb{R} \to [0,\infty)$ be a loss, $\mathcal{F} \subset \mathcal{L}_0(X)$ be a non-empty finite set, and $B > 0$ be a constant such that (6.14) holds. Then, for all measurable ERMs, all distributions P on $X \times Y$, and all $\tau > 0$, $n \geq 1$, we have*

$$\mathrm{P}^n\biggl(D \in (X\times Y)^n : \mathcal{R}_{L,\mathrm{P}}(f_D) < \mathcal{R}^*_{L,\mathrm{P},\mathcal{F}} + B\sqrt{\frac{2\tau + 2\ln(2\,|\mathcal{F}|)}{n}}\,\biggr) \;\geq\; 1 - e^{-\tau}\,.$$

Inequalities like the one above are called **oracle inequalities** since they compare the empirically obtained decision function with the one an omniscient oracle, having an infinite amount of observation, would obtain when pursuing the same goal, which in the case above is minimizing the L-risk over $\mathcal{F}$.

Proposition 6.18 shows that with high probability the function f_D approximately minimizes the risk $\mathcal{R}_{L,\mathrm{P}}$ in $\mathcal{F}$. In other words, the heuristic of replacing the unknown risk $\mathcal{R}_{L,\mathrm{P}}$ by the empirical risk $\mathcal{R}_{L,\mathrm{D}}$ is justified for *finite* sets $\mathcal{F}$. However, the assumption that $\mathcal{F}$ is finite is quite restrictive, and hence our next goal is to remove it. To this end we first observe that we cannot use a simple limit argument for $|\mathcal{F}| \to \infty$ in Proposition 6.18 since the term $B\sqrt{2\tau + 2\ln(2\,|\mathcal{F}|)}\,n^{-1/2}$ is unbounded in $|\mathcal{F}|$. To resolve this problem we introduce the following fundamental concept, which will enable us to approximate infinite $\mathcal{F}$ by *finite* subsets.

Definition 6.19. *Let (T,d) be a metric space and $\varepsilon > 0$. We call $S \subset T$ an ε-**net** of T if for all $t \in T$ there exists an $s \in S$ with $d(s,t) \leq \varepsilon$. Moreover, the ε-**covering number** of T is defined by*

$$\mathcal{N}(T,d,\varepsilon) \;:=\; \inf\biggl\{n \geq 1 : \exists\, s_1,\ldots,s_n \in T \text{ such that } T \subset \bigcup_{i=1}^n B_d(s_i,\varepsilon)\biggr\}\,,$$

where $\inf\emptyset := \infty$ and $B_d(s,\varepsilon) := \{t \in T : d(t,s) \leq \varepsilon\}$ denotes the closed ball with center $s \in T$ and radius ε.

Moreover, if (T,d) is a subspace of a normed space $(E,\|\cdot\|)$ and the metric is given by $d(x,x') = \|x - x'\|$, $x,x' \in T$, we write $\mathcal{N}(T,\|\cdot\|,\varepsilon) := \mathcal{N}(T,d,\varepsilon)$.

In simple words, an ε-net approximates T up to ε. Moreover, the covering number $\mathcal{N}(T, d, \varepsilon)$ is the size of the smallest possible ε-net, i.e., it is the smallest number of points that are needed to approximate the set T up to ε. Note that if T is compact, then all covering numbers are finite, i.e., $\mathcal{N}(T, d, \varepsilon) < \infty$ for all $\varepsilon > 0$. Moreover, $\mathcal{N}(T, d, \varepsilon)$ is a decreasing function in ε and $\sup_{\varepsilon>0} \mathcal{N}(T, d, \varepsilon) < \infty$ if and only if T is finite.

Besides covering numbers, we will also need the following "inverse" concept.

Definition 6.20. *Let (T, d) be a metric space and $n \geq 1$ be an integer. Then the n-th **(dyadic) entropy number** of (T, d) is defined by*

$$e_n(T, d) \ := \ \inf\left\{\varepsilon > 0 : \exists\, s_1, \dots, s_{2^{n-1}} \in T \text{ such that } T \subset \bigcup_{i=1}^{2^{n-1}} B_d(s_i, \varepsilon)\right\}.$$

Moreover, if (T, d) is a subspace of a normed space $(E, \|\cdot\|)$ and the metric d is given by $d(x, x') = \|x - x'\|$, $x, x' \in T$, we write

$$e_n(T, \|\cdot\|) := e_n(T, E) := e_n(T, d)\,.$$

Finally, if $S : E \to F$ is a bounded, linear operator between the normed spaces E and F, we write $e_n(S) := e_n(SB_E, \|\cdot\|_F)$.

Note that the (dyadic) entropy numbers consider ε-nets of cardinality 2^{n-1} instead of ε-nets of cardinality n. The reason for this is that this choice ensures that the entropy numbers share some basic properties with other s-numbers such as the singular numbers introduced in Section A.5.2. Basic properties of entropy numbers and their relation to singular numbers together with some bounds for important function classes can be found in Section A.5.6.

The following lemma shows that bounds on entropy numbers imply bounds on covering numbers (see Exercise 6.8 for the inverse implication).

Lemma 6.21 (Equivalence of covering and entropy numbers). *Let (T, d) be a metric space and $a > 0$ and $q > 0$ be constants such that*

$$e_n(T, d) \ \leq \ a\, n^{-1/q}\,, \qquad\qquad n \geq 1.$$

Then, for all $\varepsilon > 0$, we have

$$\ln \mathcal{N}(T, d, \varepsilon) \ \leq \ \ln(4) \cdot \left(\frac{a}{\varepsilon}\right)^q.$$

Proof. Let us fix a $\delta > 0$ and an $\varepsilon \in (0, a]$. Then there exists an integer $n \geq 1$ such that

$$a(1+\delta)(n+1)^{-1/q} \ \leq \ \varepsilon \ \leq \ a(1+\delta)n^{-1/q}\,. \tag{6.16}$$

Since $e_n(T, d) < a(1+\delta)n^{-1/q}$, there then exists an $a(1+\delta)n^{-1/q}$-net S of T with $|S| \leq 2^{n-1}$, i.e., we have

$$\mathcal{N}\bigl(T, d, a(1+\delta)n^{-1/q}\bigr) \;\le\; 2^{n-1}\,.$$

Moreover, (6.16) implies $2^{1/q}\varepsilon \ge 2^{1/q}a(1+\delta)(n+1)^{-1/q} \ge a(1+\delta)n^{-1/q}$ and $n \le \bigl(\frac{(1+\delta)a}{\varepsilon}\bigr)^q$. Consequently, we obtain

$$\ln \mathcal{N}(T, d, 2^{\frac{1}{q}}\varepsilon) \le \ln\bigl(2\mathcal{N}(T, d, a(1+\delta)n^{-\frac{1}{q}})\bigr) \le n\ln 2 \le \ln(2)\cdot\left(\frac{(1+\delta)a}{\varepsilon}\right)^q.$$

Since $\ln\mathcal{N}(T, d, \varepsilon) = 0$ for all $\varepsilon > a$, we then find the assertion. □

With the help of covering numbers, we can now investigate the statistical properties of ERM over certain *infinite* sets $\mathcal{F}$.

Proposition 6.22 (Oracle inequality for ERM). *Let* $L : X \times Y \times \mathbb{R} \to [0, \infty)$ *be a locally Lipschitz continuous loss and* P *be a distribution on* $X \times Y$. *Moreover, let* $\mathcal{F} \subset \mathcal{L}_\infty(X)$ *be non-empty and compact, and* $B > 0$ *and* $M > 0$ *be constants satisfying (6.14) and* $\|f\|_\infty \le M$, $f \in \mathcal{F}$, *respectively. Then, for all measurable ERMs and all* $\varepsilon > 0$, $\tau > 0$, *and* $n \ge 1$, *we have*

$$\mathrm{P}^n\left(\mathcal{R}_{L,\mathrm{P}}(f_D) \ge \mathcal{R}^*_{L,\mathrm{P},\mathcal{F}} + B\sqrt{\frac{2\tau + 2\ln(2\mathcal{N}(\mathcal{F}, \|\cdot\|_\infty, \varepsilon))}{n}} + 4\varepsilon|L|_{M,1}\right) \le e^{-\tau}\,.$$

Before we prove this proposition, let us first note that the compactness of $\mathcal{F}$ together with the continuity of $\mathcal{R}_{L,\mathrm{D}} : \mathcal{L}_\infty(X) \to [0, \infty)$ ensures the existence of an empirical risk minimizer. Moreover, the compactness of $\mathcal{F}$ implies that $\mathcal{F}$ is a closed and separable subset of $\mathcal{L}_\infty(X)$. The remarks after Lemma 6.17 then show that there *exists* a measurable ERM. In addition, Proposition 6.22 remains true if one replaces the compactness assumption by $\mathcal{N}(\mathcal{F}, \|\cdot\|_\infty, \varepsilon) < \infty$ for all $\varepsilon > 0$. However, in this case the *existence* of an ERM is no longer "automatically" guaranteed. Finally, if L is not locally Lipschitz continuous, variants of Proposition 6.22 still hold if the covering numbers are replaced by other notions measuring the "size" of $\mathcal{F}$. For the classification loss, corresponding results are briefly mentioned in Section 6.6.

Proof. For a fixed $\varepsilon > 0$, the compactness of $\mathcal{F}$ shows that there exists an ε-net $\mathcal{F}_\varepsilon$ of $\mathcal{F}$ with $|\mathcal{F}_\varepsilon| = \mathcal{N}(\mathcal{F}, \|\cdot\|_\infty, \varepsilon) < \infty$. For $f \in \mathcal{F}$, there thus exists a $g \in \mathcal{F}_\varepsilon$ with $\|f - g\|_\infty \le \varepsilon$, and hence we find

$$\begin{aligned}
&\bigl|\mathcal{R}_{L,\mathrm{P}}(f) - \mathcal{R}_{L,\mathrm{D}}(f)\bigr| \\
&\le \bigl|\mathcal{R}_{L,\mathrm{P}}(f) - \mathcal{R}_{L,\mathrm{P}}(g)\bigr| + \bigl|\mathcal{R}_{L,\mathrm{P}}(g) - \mathcal{R}_{L,\mathrm{D}}(g)\bigr| + \bigl|\mathcal{R}_{L,\mathrm{D}}(g) - \mathcal{R}_{L,\mathrm{D}}(f)\bigr| \\
&\le 2\,\varepsilon\,|L|_{M,1} + \bigl|\mathcal{R}_{L,\mathrm{P}}(g) - \mathcal{R}_{L,\mathrm{D}}(g)\bigr|\,,
\end{aligned}$$

where in the last step we used the local Lipschitz continuity of the L-risks established in Lemma 2.19. By taking suprema on the right- and left-hand sides we thus obtain

$$\sup_{f\in\mathcal{F}}\bigl|\mathcal{R}_{L,\mathrm{P}}(f) - \mathcal{R}_{L,\mathrm{D}}(f)\bigr| \;\le\; 2\,\varepsilon\,|L|_{M,1} + \sup_{g\in\mathcal{F}_\varepsilon}\bigl|\mathcal{R}_{L,\mathrm{P}}(g) - \mathcal{R}_{L,\mathrm{D}}(g)\bigr|\,,$$

and combining this estimate with (6.13), the latter inequality leads to

$$\begin{aligned}
&\mathrm{P}^n\biggl(D \in (X\times Y)^n : \mathcal{R}_{L,\mathrm{P}}(f_D) - \mathcal{R}^*_{L,\mathrm{P},\mathcal{F}} \geq B\sqrt{\frac{2\tau}{n}} + 4\varepsilon|L|_{M,1}\biggr)\\
&\leq \mathrm{P}^n\biggl(D \in (X\times Y)^n : \sup_{g\in\mathcal{F}_\varepsilon}\bigl|\mathcal{R}_{L,\mathrm{P}}(g) - \mathcal{R}_{L,\mathrm{D}}(g)\bigr| \geq B\sqrt{\frac{\tau}{2n}}\,\biggr)\\
&\leq 2\mathcal{N}(\mathcal{F}, \|\cdot\|_\infty, \varepsilon)\, e^{-\tau} \qquad (6.17)
\end{aligned}$$

for all $\varepsilon, \tau > 0$. Some algebraic transformations then yield the assertion. □

6.4 Basic Oracle Inequalities for SVMs

In Section 6.3, we introduced some basic techniques to analyze the statistical properties of empirical risk minimization. Since the only difference between ERM and SVMs is the additional regularization term $\lambda\|\cdot\|_H^2$, it seems plausible that these techniques can be adapted to the analysis of SVMs. This will be the idea of the second oracle inequality for SVMs we establish in this section. Moreover, we will also provide some bounds on the covering numbers for certain RKHSs. First, however, we will present another technique for establishing oracle inequalities for SVMs. This technique, which requires fewer assumptions on the kernel and the input space, combines a stability argument with the Hilbert space valued version of Hoeffding's inequality proved in Section 6.2. Finally, we illustrate how the established oracle inequalities can be used to establish both consistency and learning rates for SVMs.

Before we present the first oracle inequality, we have to ensure that SVMs are measurable learning methods. This is done in the following lemma.

Lemma 6.23 (Measurability of SVMs). *Let $L : X \times Y \times \mathbb{R} \to [0,\infty)$ be a convex loss and H be a separable RKHS over X with measurable kernel k. Then, for all $\lambda > 0$, the corresponding SVM that produces the decision functions $f_{\mathrm{D},\lambda}$ for $D \in (X\times Y)^n$ and $n \geq 1$ is a measurable learning method, and the maps $D \mapsto f_{\mathrm{D},\lambda}$ mapping $(X\times Y)^n$ to H are measurable.*

Proof. Obviously, H is a separable metric space and Lemma 4.24 ensures $H \subset \mathcal{L}_0(X)$. Moreover, the Dirac functionals are continuous on H by the definition of RKHSs, and hence the metric of H dominates the pointwise convergence. Finally, the norm $\|\cdot\|_H : H \to \mathbb{R}$ is continuous and hence measurable. Analogously to the proof of Lemma 6.17, we hence conclude that $\varphi : (X\times Y)^n \times H \to [0,\infty)$ defined by

$$\varphi(D,f) := \lambda\|f\|_H^2 + \mathcal{R}_{L,\mathrm{D}}(f)\,, \qquad D \in (X\times Y)^n,\ f \in H,$$

is measurable. In addition, Lemma 5.1 shows that $f_{\mathrm{D},\lambda}$ is the only element in H satisfying

$$\varphi(D, f_{\mathrm{D},\lambda}) = \inf_{f\in H} \varphi(D, f)\,, \qquad\qquad D \in (X \times Y)^n,$$

and consequently the measurability of $D \mapsto f_{\mathrm{D},\lambda}$ with respect to the universal completion of the product σ-algebra of $(X \times Y)^n$ follows from Aumann's measurable selection principle (see Lemma A.3.18). As in the proof of Lemma 6.17, we then obtain the first assertion. □

Let us recall that in this chapter we always assume that $(X \times Y)^n$ is equipped with the universal completion of the product σ-algebra of $(X \times Y)^n$. In addition, given a distribution P on $X \times Y$, we always write P^n for the canonical extension of the n-fold product measure of P to this completion. Note that these conventions together with Lemmas 6.23 and 6.3 make it possible to ignore measurability questions for SVMs.

Let us now establish a first oracle inequality for SVMs.

Theorem 6.24 (Oracle inequality for SVMs). *Let $L : X\times Y\times\mathbb{R} \to [0,\infty)$ be a convex, locally Lipschitz continuous loss satisfying $L(x, y, 0) \le 1$ for all $(x, y) \in X \times Y$, H be a separable RKHS over X with measurable kernel k satisfying $\|k\|_\infty \le 1$, and P be a distribution on $X \times Y$. For fixed $\lambda > 0$, $n \ge 1$, and $\tau > 0$, we then have with probability P^n not less than $1 - e^{-\tau}$ that*

$$\lambda\|f_{\mathrm{D},\lambda}\|_H^2 + \mathcal{R}_{L,\mathrm{P}}(f_{\mathrm{D},\lambda}) - \mathcal{R}^*_{L,\mathrm{P},H} < A_2(\lambda) + \lambda^{-1}|L|^2_{\lambda^{-\frac{1}{2}},1}\left(\sqrt{\frac{8\tau}{n}} + \sqrt{\frac{4}{n}} + \frac{8\tau}{3n}\right),$$

where $A_2(\,\cdot\,)$ denotes the corresponding approximation error function.

Before we prove Theorem 6.24, we note that the condition $L(x, y, 0) \le 1$ is satisfied for all margin-based losses $L(y, t) = \varphi(yt)$ for which we have $\varphi(0) \le 1$. In particular, all examples considered in Section 2.3, namely the (truncated) least squares loss, the hinge loss, and the logistic loss for classification, fall into this category. Furthermore, "restricted" distance-based losses i.e., losses $L : [-1, 1] \times \mathbb{R} \to [0,\infty)$ of the form $L(y, t) = \psi(y - t)$, $y \in [-1, 1]$, $t \in \mathbb{R}$, satisfy $L(y, 0) \le 1$, $y \in [-1, 1]$, if and only if $\psi(r) \le 1$ for all $r \in [-1, 1]$. Note that the least squares loss, the logistic loss for regression, Huber's loss for $\alpha \le \sqrt{2}$, the ϵ-insensitive loss, and the pinball loss satisfy this assumption.

Proof. Let $\Phi : X \to H$ denote the canonical feature map of k. By Corollary 5.10, there exists a bounded measurable function $h : X \times Y \to \mathbb{R}$ such that

$$\|h\|_\infty \le |L|_{\lambda^{-1/2},1}\,,$$
$$\big\|\, f_{\mathrm{P},\lambda} - f_{\mathrm{D},\lambda} \,\big\|_H \le \frac{1}{\lambda}\big\|\,\mathbb{E}_{\mathrm{P}} h\Phi - \mathbb{E}_{\mathrm{D}} h\Phi \,\big\|_H\,,$$

for all $D \in (X \times Y)^n$. Moreover, since $\|h(x, y)\Phi(x)\|_H \le \|h\|_\infty \le |L|_{\lambda^{-1/2},1}$ for all $(x, y) \in X \times Y$, we find by Corollary 6.15 that

$$\mathrm{P}^n\left(D\in(X\times Y)^n : \big\|\,\mathbb{E}_{\mathrm{P}} h\Phi - \mathbb{E}_{\mathrm{D}} h\Phi\,\big\|_H \ge |L|_{\lambda^{-1/2},1}\left(\sqrt{\frac{2\tau}{n}} + \sqrt{\frac{1}{n}} + \frac{4\tau}{3n}\right)\right) \le e^{-\tau}.$$

Combining these estimates yields

$$\mathrm{P}^n\left(D\in(X\times Y)^n : \|f_{\mathrm{D},\lambda}-f_{\mathrm{P},\lambda}\|_H \ge \lambda^{-1}|L|_{\lambda^{-\frac{1}{2}},1}\left(\sqrt{\frac{2\tau}{n}}+\sqrt{\frac{1}{n}}+\frac{4\tau}{3n}\right)\right) \le e^{-\tau}.$$

Furthermore, $\lambda\|f_{\mathrm{D},\lambda}\|_H^2 + \mathcal{R}_{L,\mathrm{D}}(f_{\mathrm{D},\lambda}) \le \lambda\|f_{\mathrm{P},\lambda}\|_H^2 + \mathcal{R}_{L,\mathrm{D}}(f_{\mathrm{P},\lambda})$ implies

$$\begin{aligned}
&\lambda\|f_{\mathrm{D},\lambda}\|_H^2 + \mathcal{R}_{L,\mathrm{P}}(f_{\mathrm{D},\lambda}) - \mathcal{R}^*_{L,\mathrm{P},H} - A_2(\lambda)\\
&= \lambda\|f_{\mathrm{D},\lambda}\|_H^2 + \mathcal{R}_{L,\mathrm{P}}(f_{\mathrm{D},\lambda}) - \lambda\|f_{\mathrm{P},\lambda}\|_H^2 - \mathcal{R}_{L,\mathrm{P}}(f_{\mathrm{P},\lambda})\\
&= \mathcal{R}_{L,\mathrm{P}}(f_{\mathrm{D},\lambda}) - \mathcal{R}_{L,\mathrm{D}}(f_{\mathrm{D},\lambda})\\
&\qquad +\lambda\|f_{\mathrm{D},\lambda}\|_H^2 + \mathcal{R}_{L,\mathrm{D}}(f_{\mathrm{D},\lambda}) - \lambda\|f_{\mathrm{P},\lambda}\|_H^2 - \mathcal{R}_{L,\mathrm{P}}(f_{\mathrm{P},\lambda})\\
&\le \mathcal{R}_{L,\mathrm{P}}(f_{\mathrm{D},\lambda}) - \mathcal{R}_{L,\mathrm{P}}(f_{\mathrm{P},\lambda}) + \mathcal{R}_{L,\mathrm{D}}(f_{\mathrm{P},\lambda}) - \mathcal{R}_{L,\mathrm{D}}(f_{\mathrm{D},\lambda})\,. \qquad (6.18)
\end{aligned}$$

Moreover, $\|f_{\mathrm{Q},\lambda}\|_\infty \le \|f_{\mathrm{Q},\lambda}\|_H \le \lambda^{-1/2}$ holds for all distributions Q on $X\times Y$ by Lemma 4.23, (5.4), and $\mathcal{R}_{L,\mathrm{Q}}(0)\le 1$. Consequently, for every distribution Q on $X\times Y$, we have

$$\mathcal{R}_{L,\mathrm{Q}}(f_{\mathrm{D},\lambda}) - \mathcal{R}_{L,\mathrm{Q}}(f_{\mathrm{P},\lambda}) \le |L|_{\lambda^{-1/2},1}\|f_{\mathrm{D},\lambda}-f_{\mathrm{P},\lambda}\|_H$$

by Lemma 2.19. Applying this estimate to (6.18) twice yields

$$\lambda\|f_{\mathrm{D},\lambda}\|_H^2 + \mathcal{R}_{L,\mathrm{P}}(f_{\mathrm{D},\lambda}) - \mathcal{R}^*_{L,\mathrm{P},H} - A_2(\lambda) \le 2|L|_{\lambda^{-1/2},1}\|f_{\mathrm{D},\lambda}-f_{\mathrm{P},\lambda}\|_H\,,$$

and by combining this inequality with the above concentration inequality we obtain the assertion. □

We will later see that a key feature of the oracle inequality above is the fact that it holds under somewhat minimal assumptions. In addition, the technique used in its proof is very flexible, as we will see, e.g., in Chapter 9 when dealing with regression problems having *unbounded* noise. On the downside, however, the oracle inequality above often leads to suboptimal learning rates. In order to illustrate this, we first need the following oracle inequality.

Theorem 6.25 (Oracle inequality for SVMs using benign kernels). *Let X be a compact metric space and $L : X\times Y\times\mathbb{R}\to[0,\infty)$ be a convex, locally Lipschitz continuous loss satisfying $L(x,y,0)\le 1$ for all $(x,y)\in X\times Y$. Moreover, let H be the RKHS of a continuous kernel k on X with $\|k\|_\infty\le 1$ and P be a probability measure on $X\times Y$. Then, for fixed $\lambda>0$, $n\ge 1$, $\varepsilon>0$, and $\tau>0$, we have with probability P^n not less than $1-e^{-\tau}$ that*

$$\begin{aligned}
&\lambda\|f_{\mathrm{D},\lambda}\|_H^2 + \mathcal{R}_{L,\mathrm{P}}(f_{\mathrm{D},\lambda}) - \mathcal{R}^*_{L,\mathrm{P},H}\\
&< A_2(\lambda) + 4\varepsilon|L|_{\lambda^{-\frac{1}{2}},1} + \bigl(|L|_{\lambda^{-\frac{1}{2}},1}\lambda^{-\frac{1}{2}}+1\bigr)\sqrt{\frac{2\tau+2\ln\bigl(2\mathcal{N}(B_H,\|\cdot\|_\infty,\lambda^{\frac{1}{2}}\varepsilon)\bigr)}{n}}.
\end{aligned}$$

Proof. By Corollary 4.31, $\mathrm{id} : H \to C(X)$ is compact, i.e., the $\|\cdot\|_\infty$-closure $\overline{B_H}$ of the unit ball B_H is a compact subset of $C(X)$. From this we conclude that $\mathcal{N}(B_H, \|\cdot\|_\infty, \varepsilon) < \infty$ for all $\varepsilon > 0$. In addition, the compactness of X implies that X is separable, and hence Lemma 4.33 shows that H is separable. Consequently, the SVM is measurable. Moreover, from (6.18) and $\|f_{\mathrm{Q},\lambda}\|_\infty \leq \|f_{\mathrm{Q},\lambda}\|_H \leq \lambda^{-1/2}$ for all distributions Q on $X \times Y$, we conclude that

$$\lambda\|f_{\mathrm{D},\lambda}\|_H^2 + \mathcal{R}_{L,\mathrm{P}}(f_{\mathrm{D},\lambda}) - \mathcal{R}^*_{L,\mathrm{P},H} - A_2(\lambda) \leq 2 \sup_{\|f\|_H \leq \lambda^{-1/2}} |\mathcal{R}_{L,\mathrm{P}}(f) - \mathcal{R}_{L,\mathrm{D}}(f)| \,.$$

In addition, for $f \in \lambda^{-1/2}B_H$ and $B := |L|_{\lambda^{-1/2},1}\lambda^{-1/2} + 1$, we have

$$\big|L(x, y, f(x))\big| \leq \big|L(x, y, f(x)) - L(x, y, 0)\big| + L(x, y, 0) \leq B$$

for all $(x, y) \in X \times Y$. Now let $\mathcal{F}_\varepsilon$ be an ε-net of $\lambda^{-1/2}B_H$ with cardinality

$$|\mathcal{F}_\varepsilon| = \mathcal{N}\big(\lambda^{-1/2}B_H, \|\cdot\|_\infty, \varepsilon\big) = \mathcal{N}\big(B_H, \|\cdot\|_\infty, \lambda^{1/2}\varepsilon\big) \,.$$

As in (6.17), we then conclude that for $\tau > 0$ we have

$$\mathrm{P}^n\left(\lambda\|f_{\mathrm{D},\lambda}\|_H^2 + \mathcal{R}_{L,\mathrm{P}}(f_{\mathrm{D},\lambda}) - \mathcal{R}^*_{L,\mathrm{P},H} \geq A_2(\lambda) + B\sqrt{\frac{2\tau}{n}} + 4\varepsilon|L|_{\lambda^{-1/2},1}\right)$$
$$\leq 2\mathcal{N}(B_H, \|\cdot\|_\infty, \lambda^{1/2}\varepsilon)\, e^{-\tau} \,.$$

By simple algebraic calculations, we then obtain the assertion. □

The right-hand side of the oracle inequality of Theorem 6.25 involves $\|\cdot\|_\infty$-covering numbers of the unit ball B_H of the RKHS H. By Corollary 4.31, these covering numbers are finite, and hence the right-hand side is non-trivial for certain values of ε, λ, and n. In order to derive consistency and learning rates from Theorem 6.25, however, we need quantitative statements on the covering numbers. This is the goal of the following two results, which for later purposes are stated in terms of entropy numbers reviewed in Section A.5.6.

Theorem 6.26 (Entropy numbers for smooth kernels). *Let $\Omega \subset \mathbb{R}^d$ be an open subset, $m \geq 1$, and k be an m-times continuously differentiable kernel on Ω. Moreover, let $X \subset \Omega$ be a closed Euclidean ball and let $H_{|X}$ denote the RKHS of the restricted kernel $k_{|X\times X}$. Assume that we have an $r_0 \in (1, \infty]$ such that $rX \subset \Omega$ for all $r \in [1, r_0)$. Then there exists a constant $c_{m,d,k}(X) > 0$ such that*

$$e_i\big(\mathrm{id} : H_{|rX} \to \ell_\infty(rX)\big) \;\leq\; c_{m,d,k}(X)\, r^m\, i^{-m/d}\,, \qquad\qquad i \geq 1, r \in [1, r_0).$$

Proof. By the definition of the space $C^0(\overline{r\mathring{X}})$ given in Section A.5.5, there exists a (unique) norm-preserving extension operator $\hat{} : C^0(\overline{r\mathring{X}}) \to C(rX)$, i.e., we have $\hat{f}_{|r\mathring{X}} = f$ and $\|f\|_\infty = \|\hat{f}\|_\infty$ for all $f \in C^0(\overline{r\mathring{X}})$. Moreover, recall that Corollary 4.36 showed that the RKHS H of k is embedded into $C^m(\Omega)$, and using the compactness of X together with (4.24) we then conclude that the restriction operator $\cdot_{|r\mathring{X}} : H_{|rX} \to C^m(\overline{r\mathring{X}})$ is continuous. Since $H_{|rX}$ consists of continuous functions, we thus obtain the commutative diagram

$$\begin{tikzcd}
H_{|rX} \arrow[r, "\mathrm{id}"] \arrow[d, "\cdot_{|r\mathring{X}}"'] & C(rX) \\
C^m(\overline{r\mathring{X}}) \arrow[r, "\mathrm{id}"'] & C^0(\overline{r\mathring{X}}) \arrow[u, "\hat{}"']
\end{tikzcd}$$

Now the multiplicity (A.38) together with (A.46), (A.47), (A.40), and the fact that $C(rX)$ is isometrically embedded into $\ell_\infty(rX)$ yields the assertion. □

Let us briefly translate the result above into the language of covering numbers. To this end, we assume that X and k satisfy the assumptions of Theorem 6.26. Lemma 6.21 then shows that

$$\ln \mathcal{N}(B_{H_{|rX}}, \|\cdot\|_\infty, \varepsilon) \;\leq\; a\,\varepsilon^{-2p}\,, \qquad\qquad \varepsilon > 0. \tag{6.19}$$

for $2p := d/m$ and $a := \ln(4) \cdot (c_{m,d}(X))^{d/m}\, r^d$. Now recall that Taylor and Gaussian RBF kernels are infinitely often differentiable and hence (6.19) holds for arbitrarily small $p > 0$. For Gaussian RBF kernels, however, the parameter γ is usually *not* fixed (see Section 8.2), and hence it is important to know how the constant a depends on γ. This is the goal of the next theorem.

Theorem 6.27 (Entropy numbers for Gaussian kernels). *Let $X \subset \mathbb{R}^d$ be a closed Euclidean ball and $m \geq 1$ be an integer. Then there exists a constant $c_{m,d}(X) > 0$ such that, for all $0 < \gamma \leq r$ and all $i \geq 1$, we have*

$$e_i\big(\mathrm{id} : H_\gamma(rX) \to \ell_\infty(rX)\big) \;\leq\; c_{m,d}(X)\, r^m\, \gamma^{-m} i^{-\frac{m}{d}}\,.$$

Proof. For $x \in r\gamma^{-1}X$ and $f \in H_\gamma(rX)$, we write $\tau_\gamma f(x) := f(\gamma x)$. Proposition 4.37 applied to the dilation factor γ, the kernel parameter 1, and the set $r\gamma^{-1}X$ then shows that $\tau_\gamma : H_\gamma(rX) \to H_1(r\gamma^{-1}X)$ is an isometric isomorphism. Moreover, the dilation $\tau_{1/\gamma} : \ell_\infty(r\gamma^{-1}X) \to \ell_\infty(rX)$ is clearly an isometric isomorphism, too. In addition, we have the commutative diagram

$$\begin{tikzcd}
H_\gamma(rX) \arrow[r, "\mathrm{id}"] \arrow[d, "\tau_\gamma"'] & \ell_\infty(rX) \\
H_1(r\gamma^{-1}X) \arrow[r, "\mathrm{id}"'] & \ell_\infty(r\gamma^{-1}X) \arrow[u, "\tau_{1/\gamma}"']
\end{tikzcd}$$

and hence we obtain the assertion by Theorem 6.26 and (A.38). □

Our last goal in this section is to illustrate how the above oracle inequalities can be used to establish both consistency and learning rates for SVMs. For conceptional simplicity, we thereby restrict our considerations to *Lipschitz continuous* losses L with $|L|_1 \leq 1$, but similar results can be easily derived for locally Lipschitz continuous losses, too. Now recall that the Lipschitz continuity together with $L(x,y,0) \leq 1$ yields $L(x,y,t) \leq 1 + |t|$, and hence L is a P-integrable Nemitski loss of order 1 for all distributions P on $X \times Y$. In the following we further assume for simplicity that we use a *fixed* RKHS H that in addition is assumed to be dense in $L_1(\mu)$ for all distributions μ on X. Here we recall that we have intensively investigated such RKHSs in Section 4.6. Moreover, Theorem 5.31 showed for such H that $\mathcal{R}^*_{L,\mathrm{P},H} = \mathcal{R}^*_{L,\mathrm{P}}$, i.e., the Bayes risk can be approximated by functions from H.

In the following, we only consider the situation of Theorem 6.25 since for Theorem 6.24 the results are similar (see Exercise 6.9 for precise statements and a comparison of the resulting learning rates). Since Theorem 6.25 involves covering numbers, we assume for simplicity that there exist constants $a \geq 1$ and $p > 0$ such that

$$\ln \mathcal{N}(B_H, \|\cdot\|_\infty, \varepsilon) \;\leq\; a\,\varepsilon^{-2p}\,, \qquad\qquad \varepsilon > 0. \tag{6.20}$$

By Theorem 6.26 and Lemma 6.21, we see that both Taylor and Gaussian kernels satisfy this assumption for all $p > 0$. Moreover, we saw in Section 4.6 that *(a)* Taylor kernels often have RKHSs that are dense in $L_1(\mu)$ and *(b)* Gaussian kernels always satisfy this denseness assumption. Consequently, these kernels are ideal candidates for our discussion.

In order to illustrate the utility of the oracle inequalities obtained let us now fix a $\lambda \in (0,1]$ and a $\tau \geq 1$. For

$$\varepsilon := \Bigl(\frac{p}{2}\Bigr)^{1/(1+p)} \Bigl(\frac{2a}{n}\Bigr)^{1/(2+2p)} \lambda^{-1/2}\,.$$

Theorem 6.25 together with Lemma A.1.5 and $(p+1)(2/p)^{p/(1+p)} \leq 3$ then shows that

$$\lambda\|f_{\mathrm{D},\lambda}\|_H^2 + \mathcal{R}_{L,\mathrm{P}}(f_{\mathrm{D},\lambda}) - \mathcal{R}^*_{L,\mathrm{P}} < A_2(\lambda) + \frac{3}{\lambda^{1/2}}\biggl(2\Bigl(\frac{2a}{n}\Bigr)^{\frac{1}{2+2p}} + \Bigl(\frac{2\tau}{n}\Bigr)^{\frac{1}{2}}\biggr) \tag{6.21}$$

holds with probability P^n not less than $1 - e^{-\tau}$.

Let us now assume that for sample size n we choose a $\lambda_n \in (0,1]$ such that $\lim_{n\to\infty} \lambda_n = 0$ and

$$\lim_{n\to\infty} \lambda_n^{1+p} n = \infty\,. \tag{6.22}$$

Lemma 5.15, see also (5.32), then shows that the right-hand side of (6.21) converges to 0, and hence we have $\mathcal{R}_{L,\mathrm{P}}(f_{\mathrm{D},\lambda_n}) \to \mathcal{R}^*_{L,\mathrm{P}}$ in probability. In other words, we have shown that, for RKHSs satisfying both the denseness assumption above and (6.20), the SVM is universally L-risk consistent whenever the regularization sequence tends to zero in a controlled way described by (6.22).

In order to establish learning rates, let us additionally assume that there exist constants $c > 0$ and $\beta \in (0, 1]$ such that

$$A_2(\lambda) \le c\lambda^\beta\,, \qquad \lambda \ge 0. \tag{6.23}$$

Then a straightforward calculation shows that the asymptotically best choice for λ_n in (6.21) is a sequence that behaves like $n^{-\frac{1}{(1+p)(2\beta+1)}}$ and that the resulting learning rate is given by

$$\mathrm{P}^n\Bigl(D \in (X \times Y)^n : \mathcal{R}_{L,\mathrm{P}}(f_{\mathrm{D},\lambda_n}) - \mathcal{R}^*_{L,\mathrm{P}} \le C\sqrt{\tau}\, n^{-\frac{\beta}{(2\beta+1)(1+p)}}\Bigr) \ge 1 - e^{-\tau}\,,$$

where C is a constant independent of τ and n. It is important to note that the regularization sequence (λ_n) that achieves this rate *depends* on β. Unfortunately, however, we will almost never know the value of β, and hence we cannot choose the "optimal" regularization sequence suggested by Theorem 6.25. In the following section, we will therefore investigate how this problem can be addressed by choosing λ in a data-dependent way.

6.5 Data-Dependent Parameter Selection for SVMs

In this section, we first present a simple method for choosing the regularization parameter λ in a *data-dependent* way. We will then show that this method is *adaptive* in the sense that it does not need to know characteristics of the distribution such as (6.23) to achieve the learning rates we obtained in the previous section by knowing these characteristics.

Let us begin by describing this parameter selection method, which in some sense is a simplification of cross-validation considered in Section 11.3.

Definition 6.28. *Let $L : X \times Y \times \mathbb{R} \to [0, \infty)$ be a convex loss that can be clipped at 1, H be an RKHS over X, and $\Lambda := (\Lambda_n)$ be a sequence of finite subsets $\Lambda_n \subset (0, 1]$. Given a $D := ((x_1, y_1), \ldots, (x_n, y_n)) \in (X \times Y)^n$, we define*

$$D_1 := ((x_1, y_1), \ldots, (x_m, y_m))\,,$$
$$D_2 := ((x_{m+1}, y_{m+1}), \ldots, (x_n, y_n))\,,$$

where $m := \lfloor n/2 \rfloor + 1$ and $n \ge 3$. Then use D_1 as a training set by computing the SVM decision functions

$$f_{\mathrm{D}_1,\lambda} := \arg\min_{f \in H} \lambda\|f\|_H^2 + \mathcal{R}_{L,\mathrm{D}_1}(f)\,, \qquad \lambda \in \Lambda_n, \tag{6.24}$$

and use D_2 to determine λ by choosing a $\lambda_{D_2} \in \Lambda_n$ such that

$$\mathcal{R}_{L,\mathrm{D}_2}(\widehat{f}_{\mathrm{D}_1,\lambda_{D_2}}) = \min_{\lambda \in \Lambda_n} \mathcal{R}_{L,\mathrm{D}_2}(\widehat{f}_{\mathrm{D}_1,\lambda})\,, \tag{6.25}$$

where $\widehat{f}_{\mathrm{D}_1,\lambda}$ denotes the clipped version of $f_{\mathrm{D}_1,\lambda}$. Every learning method that produces the resulting decision functions $f_{\mathrm{D}_1,\lambda_{D_2}}$ is called a ***training validation support vector machine (TV-SVM)*** *with respect to Λ.*

Informally speaking, the idea of TV-SVMs[2] is to use the *training set* D_1 to build a couple of SVM decision functions and then use the decision function that best performs on the independent *validation set* D_2. Here we note that Theorem 5.5 ensures that the SVM solutions $f_{\mathrm{D}_1,\lambda}$, $\lambda \in \Lambda_n$, found in the training step (6.24) exist, and hence there exists a TV-SVM. However, note that in general the validation step (6.25) does *not* provide a unique regularization parameter λ_{D_2}, and hence the TV-SVM, like ERM, is not a uniquely defined learning method. The following lemma shows that for all interesting cases there exists a measurable TV-SVM.

Lemma 6.29 (Measurability of TV-SVMs). *Let $L : X \times Y \times \mathbb{R} \to [0,\infty)$ be a convex loss that can be clipped at 1, and let H be a separable RKHS over X having a measurable kernel. Then there exists a measurable TV-SVM.*

Proof. Lemma 6.23 showed that $(D,x) \mapsto f_{\mathrm{D}_1,\lambda}(x)$ is measurable, and hence $\varphi : (X \times Y)^n \times \Lambda_n \to [0,\infty)$ defined by

$$\varphi(D,\lambda) := \mathcal{R}_{L,\mathrm{D}_2}(\widehat{f}_{\mathrm{D}_1,\lambda}), \qquad\qquad D \in (X\times Y)^n,\ \lambda \in \Lambda_n,$$

is measurable. The rest of the proof is analogous to the proofs of Lemmas 6.17 and 6.23. □

Our next goal is to establish oracle inequalities for TV-SVMs. To this end, we need the following lemma that describes how the term on the right-hand side of our oracle inequalities for SVMs can be approximately minimized.

Lemma 6.30. *Let $L : X \times Y \times \mathbb{R} \to [0,\infty)$ be a loss, H be the RKHS of a measurable kernel over X, P be a distribution on $X \times Y$ with $\mathcal{R}^*_{L,\mathrm{P},H} < \infty$, and $A_2 : [0,\infty) \to [0,\infty)$ be the corresponding approximation error function. We fix a bounded interval $I \subset (0,\infty)$. In addition, let $\alpha, c \in (0,\infty)$ be two constants and Λ be a finite ε-net of I for some fixed $\varepsilon > 0$. Then we have*

$$\min_{\lambda\in\Lambda}\bigl(A_2(\lambda) + c\lambda^{-\alpha}\bigr) \;\le\; A_2(2\varepsilon) + \inf_{\lambda\in I}\bigl(A_2(\lambda) + c\lambda^{-\alpha}\bigr)\,.$$

Proof. Let us assume that Λ is of the form $\Lambda = \{\lambda_1,\dots,\lambda_m\}$ with $\lambda_{i-1} < \lambda_i$ for all $i = 2,\dots,m$. We write $\lambda_0 := \inf I$. Our first goal is to show that

$$\lambda_i - \lambda_{i-1} \;\le\; 2\varepsilon\,, \qquad\qquad i = 1,\dots,m. \tag{6.26}$$

To this end, we fix an $i \in \{1,\dots,m\}$ and write $\bar\lambda := (\lambda_i + \lambda_{i-1})/2 \in I \cup \{\lambda_0\}$. Since $\Lambda \cup \{\lambda_0\}$ is an ε-net of $I \cup \{\lambda_0\}$, we then have $\lambda_i - \bar\lambda \le \varepsilon$ or $\bar\lambda - \lambda_{i-1} \le \varepsilon$. Simple algebra shows that in both cases we find (6.26). For $\delta > 0$, we now fix a $\lambda^* \in I$ such that

$$A_2(\lambda^*) + c(\lambda^*)^{-\alpha} \le \inf_{\lambda\in I}\bigl(A_2(\lambda) + c\lambda^{-\alpha}\bigr) + \delta\,. \tag{6.27}$$

[2] For simplicity, we only consider (almost) equally sized data sets D_1 and D_2, but the following results and their proofs remain almost identical for different splits.

Then there exists an index $i \in \{1, \dots, m\}$ such that $\lambda_{i-1} \leq \lambda^* \leq \lambda_i$, and by (6.26) we conclude that $\lambda^* \leq \lambda_i \leq \lambda^* + 2\varepsilon$. By the monotonicity and subadditivity of $A_2(\,\cdot\,)$ established in Lemma 5.15, we thus find

$$\begin{aligned}\min_{\lambda \in \varLambda}\bigl(A_2(\lambda) + c\lambda^{-\alpha}\bigr) \leq A_2(\lambda_i) + c\lambda_i^{-\alpha} &\leq A_2(\lambda^* + 2\varepsilon) + c(\lambda^*)^{-\alpha} \\ &\leq A_2(\lambda^*) + c(\lambda^*)^{-\alpha} + A_2(2\varepsilon)\,.\end{aligned}$$

Combining this estimate with (6.27) then yields the assertion. □

With the help of the Lemma 6.30, we can now establish our first oracle inequality for TV-SVMs. For simplicity, it only considers the situation investigated at the end of Section 6.4, but generalizations are easy to establish.

Theorem 6.31 (Oracle inequality for TV-SVMs and benign kernels). *Let X be a compact metric space and $L : X \times Y \times \mathbb{R} \to [0, \infty)$ be a convex, Lipschitz continuous loss with $|L|_1 \leq 1$. Assume that L can be clipped at 1 and that it satisfies $L(x, y, 0) \leq 1$ for all $(x, y) \in X \times Y$. Furthermore, let H be the RKHS of a continuous kernel k on X satisfying $\|k\|_\infty \leq 1$ and*

$$\ln \mathcal{N}(B_H, \|\cdot\|_\infty, \varepsilon) \;\leq\; a\varepsilon^{-2p}\,, \qquad\qquad \varepsilon > 0, \tag{6.28}$$

where $a \geq 1$ and $p > 0$ are constants. Moreover, for $n \geq 4$ and $\varepsilon > 0$, let $\varLambda_n \subset (0, 1]$ be a finite ε-net of $(0, 1]$ of cardinality $|\varLambda_n|$. For fixed $\tau > 0$ and $\tau_n := 2 + \tau + \ln|\varLambda_n|$, we then have with probability P^n not less than $1 - e^{-\tau}$ that

$$\mathcal{R}_{L,\mathrm{P}}(\widehat{f}_{\mathrm{D}_1, \lambda_{D_2}}) - \mathcal{R}^*_{L,\mathrm{P},H} < \inf_{\lambda \in (0,1]} \biggl(A_2(\lambda) + \frac{13}{\sqrt{\lambda}} \biggl(\Bigl(\frac{a}{n}\Bigr)^{\frac{1}{2+2p}} + \sqrt{\frac{\tau_n}{n}}\,\biggr)\biggr) + A_2(2\varepsilon)\,.$$

Consequently, if we use ε_n-nets $\varLambda_n$ with $\varepsilon_n \to 0$ and $n^{-1} \ln|\varLambda_n| \to 0$, then the resulting TV-SVM is consistent for all P satisfying $\mathcal{R}^*_{L,\mathrm{P},H} = \mathcal{R}^*_{L,\mathrm{P}}$. Finally, if $\varepsilon_n \leq 1/n$ and $|\varLambda_n|$ grows polynomially in n, then the TV-SVM learns with rate

$$n^{-\frac{\beta}{(2\beta+1)(1+p)}} \tag{6.29}$$

for all distributions P that satisfy $A_2(\lambda) \leq c\lambda^\beta$ for some constants $c > 0$ and $\beta \in (0, 1]$ and all $\lambda \geq 0$.

Proof. Let us define $m := \lfloor n/2 \rfloor + 1$. Since $m \geq n/2$, we obtain similarly to (6.21) that with probability P^m not less than $1 - |\varLambda_n|e^{-\tau}$ we have

$$\mathcal{R}_{L,\mathrm{P}}(f_{\mathrm{D}_1,\lambda}) - \mathcal{R}^*_{L,\mathrm{P},H} < A_2(\lambda) + \frac{3}{\lambda^{1/2}}\biggl(2\Bigl(\frac{4a}{n}\Bigr)^{\frac{1}{2+2p}} + \Bigl(\frac{2\tau + 2}{n}\Bigr)^{\frac{1}{2}}\biggr)$$

for all $\lambda \in \varLambda_n$ *simultaneously.* In addition, we have $L(x, y, \widehat{t}) \leq |L|_1 + L(x, y, 0) \leq 2 =: B$, and hence Proposition 6.18 yields

$$\mathrm{P}^{n-m}\biggl(D_2 : \mathcal{R}_{L,\mathrm{P}}(\wideparen{f}_{\mathrm{D}_1,\lambda_{D_2}}) < \inf_{\lambda\in\Lambda_n} \mathcal{R}_{L,\mathrm{P}}(\wideparen{f}_{\mathrm{D}_1,\lambda}) + 4\sqrt{\frac{2\tau+2\ln(2|\Lambda_n|)}{n}}\biggr) \geq 1-e^{-\tau},$$

where we used $n-m \geq n/2-1 \geq n/4$. Since $\mathcal{R}_{L,\mathrm{P}}(\wideparen{f}_{\mathrm{D}_1,\lambda}) \leq \mathcal{R}_{L,\mathrm{P}}(f_{\mathrm{D}_1,\lambda})$, we conclude that with probability P^n not less than $1-(|\Lambda_n|+1)e^{-\tau}$ we have

$$\begin{aligned}
\mathcal{R}_{L,\mathrm{P}}(\wideparen{f}_{\mathrm{D}_1,\lambda_{D_2}}) - \mathcal{R}^*_{L,\mathrm{P},H} &< \inf_{\lambda\in\Lambda_n}\biggl(A_2(\lambda) + \frac{3}{\lambda^{1/2}}\Bigl(2\Bigl(\frac{4a}{n}\Bigr)^{\frac{1}{2+2p}} + \Bigl(\frac{2\tau+2}{n}\Bigr)^{\frac{1}{2}}\Bigr)\biggr) \\
&\qquad + 4\sqrt{\frac{2\tau+2\ln(2|\Lambda_n|)}{n}} \\
&\leq \inf_{\lambda\in(0,1]}\biggl(A_2(\lambda) + \frac{3}{\lambda^{1/2}}\Bigl(2\Bigl(\frac{4a}{n}\Bigr)^{\frac{1}{2+2p}} + \Bigl(\frac{2\tau+2}{n}\Bigr)^{\frac{1}{2}}\Bigr)\biggr) \\
&\qquad + A_2(2\varepsilon) + 4\sqrt{\frac{2\tau+2\ln(2|\Lambda_n|)}{n}}\,,
\end{aligned}$$

where in the last step we used Lemma 6.30. From this we easily obtain the first assertion. The second and third assertions then follow by the arguments used at the end of Section 6.4. □

Note that the preceding proof heavily relied on the assumption that L can be clipped. Indeed, without this assumption, Proposition 6.18 only shows that

$$\mathcal{R}_{L,\mathrm{P}}(f_{\mathrm{D}_1,\lambda_{D_2}}) < \inf_{\lambda\in\Lambda_n} \mathcal{R}_{L,\mathrm{P}}(f_{\mathrm{D}_1,\lambda}) + 4 \sup_{\lambda\in\Lambda_n} \lambda^{-1/2}\sqrt{\frac{2\tau+2\ln(2|\Lambda_n|)}{n}}$$

holds with probability not less than $1-e^{-\tau}$. Since for n^{-1}-nets Λ_n of $(0,1]$ we have $\sup_{\lambda\in\Lambda_n} \lambda^{-1/2} \geq n^{1/2}$, it becomes obvious that the preceding proof does not provide consistency or the rates (6.29) if L cannot be clipped. In other words, the fact that L is clippable ensures that the error of the parameter selection step does *not* dominate the error of the SVM training step.

Let us now recall the end of Section 6.4, where we saw that SVMs satisfying the covering number assumption (6.28) and the approximation error assumption $A_2(\lambda) \leq c\lambda^\beta$ can learn with rate (6.29). Unfortunately, however, this rate required a regularization sequence $\lambda_n := n^{-\frac{1}{(1+p)(2\beta+1)}}$, i.e., the rate was only achievable if we had knowledge on the distribution P, the RKHS H, and their interplay. Of course, we almost never know the exponent β that bounds the approximation error function, and hence it remained unclear whether the learning rate (6.29) was actually realizable. Theorem 6.31 now shows that the TV-SVM does achieve this learning rate *without* knowing the exponent β. Moreover, the theorem also shows that we do not even have to know the exponent p in the covering number assumption (6.28) to achieve this rate. Of course, this p is independent of P and hence in principle *a priori* known. In practice, however, covering number bounds are often extremely difficult to establish for new RKHSs, and hence the independence of the TV-SVM from this exponent is an important feature. Furthermore, the following oracle inequality for TV-SVMs shows that this learning method achieves non-trivial learning rates even if there is no exponent p satisfying (6.28).

Theorem 6.32 (Oracle inequality for TV-SVMs). *Let $L : X \times Y \times \mathbb{R} \to [0, \infty)$ be a convex Lipschitz continuous loss that can be clipped at 1 and that satisfies $|L|_1 \leq 1$ and $L(x, y, 0) \leq 1$ for all $(x, y) \in X \times Y$. Furthermore, let H be a separable RKHS with measurable kernel k over X satisfying $\|k\|_\infty \leq 1$. Moreover, for $n \geq 4$ and $\varepsilon > 0$, let $\Lambda_n \subset (0, 1]$ be a finite ε-net of $(0, 1]$. For fixed $\tau > 0$, we then have with probability P^n not less than $1 - e^{-\tau}$ that*

$$\mathcal{R}_{L,\mathrm{P}}(\widehat{f}_{\mathrm{D}_1,\lambda_{D_2}}) - \mathcal{R}^*_{L,\mathrm{P},H} < \inf_{\lambda\in(0,1]} \biggl(A_2(\lambda) + \frac{14}{\lambda} \biggl(\sqrt{\frac{\tau + \ln(2|\Lambda_n|)}{n}} + \frac{\tau + \ln(2|\Lambda_n|)}{n} \biggr) \biggr) \\ + A_2(2\varepsilon)\,.$$

In particular, if we use ε_n-nets Λ_n with $\varepsilon_n \to 0$ and $n^{-1} \ln |\Lambda_n| \to 0$, then the resulting TV-SVM is consistent for all P with $\mathcal{R}^*_{L,\mathrm{P},H} = \mathcal{R}^*_{L,\mathrm{P}}$. Moreover, for $\varepsilon_n \leq n^{-1/2}$ and $|\Lambda_n|$ growing polynomially in n, the TV-SVM learns with rate

$$\Bigl(\frac{\ln(n + 1)}{n} \Bigr)^{\frac{\beta}{2\beta+2}} \tag{6.30}$$

for all distributions P that satisfy $A_2(\lambda) \leq c\lambda^\beta$ for some constants $c > 0$ and $\beta \in (0, 1]$ and all $\lambda \geq 0$.

Proof. Repeat the proof of Theorem 6.31, but use Theorem 6.24 instead of Theorem 6.25. □

Theorem 6.32 shows that the TV-SVM learns with a specific rate if an approximation error assumption is satisfied. Moreover, this rate equals the "optimal" rate we can derive from Theorem 6.24 up to a logarithmic factor (see Exercise 6.9), i.e., the TV-SVM is again adaptive with respect to the unknown exponent β bounding the approximation error function. Moreover, by combining the two oracle inequalities for the TV-SVM, we see that the TV-SVM is in some sense also adaptive to the size of the input domain. To illustrate this, let us consider the space $X := \mathbb{R}^d$. Moreover, assume that we have an RKHS H over X such that the covering number bound (6.28) is satisfied for some exponent $p(X')$ whenever we consider the restriction of H to some compact subset $X' \subset \mathbb{R}^d$. By combining Theorem 6.31 with Theorem 6.32, we then see that the TV-SVM learns with rate (6.30) if the support of P_X is not compact and with rate

$$\min \biggl\{ \Bigl(\frac{\ln(n + 1)}{n} \Bigr)^{\frac{\beta}{2\beta+2}}, n^{-\frac{\beta}{(1+p(X'))(2\beta+1)}} \biggr\}$$

if $X' := \operatorname{supp}(\mathrm{P}_X)$ is compact. In this sense, the TV-SVM is adaptive not only to the approximation error assumption (6.23) but also to the input domain of the data. Finally, note that these considerations can be refined using the more advanced techniques of the next chapter. We refer to Section 8.3, where this is worked out in detail for binary classification.

6.6 Further Reading and Advanced Topics

The first learning method that was shown to be universally consistent (see Stone, 1977) was the so-called nearest-neighbor method. Since then, universal consistency has been established for a variety of different methods. Many examples of such methods for classification and regression can be found in the books by Devroye *et al.* (1996) and Györfi *et al.* (2002), respectively. Moreover, besides the no-free-lunch theorem, which was proved by Devroye (1982), Devroye *et al.* (1996) also present some other fundamental limitations in statistical learning theory. These limitations include the non-existence of an overall best-performing classification method, the no-free-lunch theorem under certain additional assumptions on P, and the non-existence of a method that estimates the Bayes risk with a uniform rate. Moreover, learning rates (and their optimality) for certain regression methods are presented in great detail by Györfi *et al.* (2002).

The classical concentration inequalities presented in Section 6.2 were proven by Hoeffding (1963) and Bernstein (1946). Sharper versions of Bernstein's inequality were found by Bennett (1962) and Hoeffding (1963). For a more detailed discussion on these inequalities, we refer to Hoeffding (1963) and Bousquet (2003a). Finally, Theorem 6.13 and the Hilbert space valued versions of Bernstein's and Hoeffding's inequalities were taken from Chapter 3 of Yurinsky (1995). Note that the crucial step in deriving these Hilbert space valued versions is the estimate (6.10), which by symmetrization holds (up to some constant) in every Banach space of type 2. Moreover, weaker versions of (6.10) can actually be established whenever the Banach space has some non-trivial type. For more information on the type concept for Banach spaces, we refer to Chapter 11 of Diestel *et al.* (1995).

The discussion in Section 6.3 is nowadays folklore in the machine learning literature. The idea of estimating the excess risk of an empirical risk minimizer by a supremum (6.13) goes back to Vapnik and Chervonenkis (1974). Generalizations of this bound to infinite sets $\mathcal{F}$ require bounds on the "size" or "complexity" of $\mathcal{F}$. Probably the most classical such complexity measure is the so-called Vapnik-Chervonenkis (VC) dimension, which can be applied if, e.g., L is the binary classification loss. Furthermore, there are various extensions and generalizations of the VC dimension that make it possible to deal with other types of loss functions. We refer to the books by Vapnik (1998), Anthony and Bartlett (1999), and Vidyasagar (2002).

Using covering numbers as a complexity measure is another idea that frequently appears in the literature. Probably the easiest way to use these numbers is presented in (6.17), but there also exist more sophisticated concentration inequalities, such as Lemma 3.4 of Alon *et al.* (1997) and Theorem 9.1 of Györfi *et al.* (2002), where the latter goes back to Pollard (1984). Covering numbers themselves were first investigated by Kolmogorov (1956) and Kolmogorov and Tikhomirov (1961). Since then various results for interesting function classes have been established. We refer to the books of Pinkus (1985), Carl and Stephani (1990), and Edmunds and Triebel (1996) for a detailed account and to Section A.5.6 for a brief overview.

Results similar to Theorem 6.25 were first established by Cucker and Smale (2002) and Steinwart (2005). Moreover, results in the spirit of Theorem 6.24 were found by Zhang (2001), Steinwart (2005), and in a different context by Bousquet and Elisseeff (2002). Universal consistency of SVMs for binary classification was first shown by Steinwart (2002), Zhang (2004b), and Steinwart (2005). Finally, consistency of SVMs for certain violations of the i.i.d. assumption was recently shown by Steinwart *et al.* (2008) and Steinwart and Anghel (2008) with techniques similar to the one used for Theorem 6.24.

In its simplistic form, the parameter selection method considered in Section 6.5 is little more than an illustration of how oracle inequalities can be used to analyze learning methods that *include* the parameter selection step. Nonetheless, the TV-SVM procedure is related to commonly used methods such as grid search and cross-validation, discussed in Section 11.3. A different approach for the parameter selection problem is considered by Lecué (2007b), who proposes to use the aggregated decision function

$$\sum_{\lambda\in\Lambda} w_\lambda \widehat{f}_{\mathrm{D}_1,\lambda}\,,$$

where the weights w_λ are computed in terms of $\mathcal{R}_{L,\mathrm{D}_2}(\widehat{f}_{\mathrm{D}_1,\lambda})$. More precisely, he considers weights of the form

$$w_\lambda := \frac{\exp(-|D_2|\,\mathcal{R}_{L,\mathrm{D}_2}(\widehat{f}_{\mathrm{D}_1,\lambda}))}{\sum_{\lambda'\in\Lambda}\exp(-|D_2|\,\mathcal{R}_{L,\mathrm{D}_2}(\widehat{f}_{\mathrm{D}_1,\lambda'}))}$$

and establishes, for example for the hinge loss, oracle inequalities for this approach. These oracle inequalities imply that this aggregation procedure is adaptive to characteristics of P considered in Chapter 8. Moreover, a similar weighting approach was taken by Bunea and Nobel (2005) for the least squares loss. For further methods and results, we refer to Bartlett (2008), Bunea *et al.* (2007), Dalalyan and Tsybakov (2007), Lecué (2007a), Tsybakov (2003), and the references therein.

6.7 Summary

In this chapter, we developed basic techniques for investigating the statistical properties of SVMs. To this end, we first introduced two notions of statistical learning, namely the purely asymptotic notion of *consistency* and the more practically oriented notion of *learning rates*. We further presented the *no-free-lunch theorem*, which implied that uniform learning rates are impossible without assumptions on P.

In Section 6.2, we then established *concentration inequalities*, which described how close empirical averages of i.i.d. random variables are centered around their mean. The main results in this direction were *Hoeffding's inequality*, which gives an exponential tail for bounded real-valued random variables, and *Bernstein's inequality*, which improves this tail when the variance of the random variables is substantially smaller than their supremum norm. Finally, we generalized these inequalities to Hilbert space valued random variables.

In Section 6.3, we used these inequalities to analyze empirical risk minimization. We began by considering empirical risk minimizers over *finite* function classes and introduced *covering* and *entropy numbers* to generalize the basic idea to infinite function classes. The techniques developed for ERM were then modified in Section 6.4 to establish *oracle inequalities* for SVMs. There we also illustrated how these oracle inequalities can be used to establish both consistency and learning rates for SVMs whose regularization parameter only depends on the sample size. Unfortunately, however, the fastest learning rates we obtained required knowledge about certain characteristics of the data-generating distribution P. Since this knowledge is typically not available, we finally introduced and analyzed a data-dependent choice of the regularization parameter in Section 6.5. This selection method turned out to be consistent and, more important, we also saw that this method is *adaptive* to some unknown characteristics of P.

6.8 Exercises

6.1. Comparison of Hoeffding's and Bernstein's inequalities ($\star$)
Let $(\Omega, \mathcal{A}, \mathrm{P})$ be a probability space, $B > 0$, and $\sigma > 0$. Furthermore, let $\xi_1, \dots, \xi_n : \Omega \to \mathbb{R}$ be independent and bounded random variables with $\|\xi_i\|_\infty \leq B$ and $\mathbb{E}\xi_i^2 \leq \sigma^2$ for all $i = 1, \dots, n$. Finally, let $\tau > 0$ be a real number and $n \geq 1$ be an integer satisfying $n \geq \frac{8}{9}\tau$. Show that Bernstein's inequality is sharper than Hoeffding's inequality if and only if

$$\sigma < \left(1 - \sqrt{\frac{8\tau}{9n}}\right) B .$$

What happens if we additionally assume $\mathbb{E}\xi_i = 0$ for all $i = 1, \dots, n$?

6.2. A variant of Markov's inequality ($\star\star$)
Let $(\Omega, \mathcal{A}, \mathrm{P})$ be a probability space and $f : \Omega \to \mathbb{R}$ be a measurable function. Show that for all $t > 0$ the following inequalities hold:

$$\sum_{n=1}^{\infty} \mathrm{P}\bigl(\{\omega \in \Omega : |f(\omega)| \geq nt\}\bigr) \leq \frac{\mathbb{E}_{\mathrm{P}}|f|}{t} \leq 1 + \sum_{n=1}^{\infty} \mathrm{P}\bigl(\{\omega \in \Omega : |f(\omega)| \geq nt\}\bigr).$$

Hint: Apply Lemma A.3.11.

6.3. Chebyshev's inequality for sums of i.i.d. random variables ($\star\star$)
Let $(\Omega, \mathcal{A}, \mathrm{P})$ be a probability space and $\xi_1, \dots, \xi_n : \Omega \to \mathbb{R}$ be independent random variables for which there exists a constant $\sigma > 0$ such that $\mathbb{E}_{\mathrm{P}}\xi_i^2 \leq \sigma^2$ for all $i = 1, \dots, n$.
i). Show the following inequality:

$$\mathrm{P}\biggl(\frac{1}{n}\sum_{i=1}^{n}(\xi_i - \mathbb{E}\xi_i) \geq \sqrt{\frac{2\sigma^2 e^{\tau}}{n}}\biggr) \leq e^{-\tau}\,, \qquad \tau > 0.$$

ii). Compare this inequality with Hoeffding's and Bernstein's inequalities.
iii). Generalize the inequality above to Hilbert space valued random variables.

6.4. Proof of the no-free-lunch theorem($\star\star\star\star$)
Prove Theorem 6.6 using the proof of Theorem 7.2 by Devroye *et al.* (1996).

Hint: Fix an arbitrary decreasing sequence $(p_i) \subset (0, 1]$ with $\sum p_i = 1$. Using Lyapunov's Theorem A.3.13, which in particular states that $\{\mu(A) : A \in \mathcal{A}\} = [0, 1]$, construct a sequence (A_i) of mutually disjoint $A_i \in \mathcal{A}$ satisfying $\mu(A_i) = p_i$ for all $i \geq 1$. Use this to suitably modify the construction at the beginning of the proof of Theorem 7.2 by Devroye *et al.* (1996). Check that the rest of the proof can be kept unchanged.

6.5. No uniform rate for convex losses ($\star\star\star$)
Let $L : X \times Y \times \mathbb{R} \to [0, \infty)$ be a convex loss function for which there exist two distributions Q_1 and Q_2 on Y that have mutually distinct L-risk minimizers, i.e., for all $x \in X$, we have $\mathcal{M}_{L,\mathrm{Q}_1,x}(0^+) \neq \emptyset$, $\mathcal{M}_{L,\mathrm{Q}_2,x}(0^+) \neq \emptyset$, and

$$\mathcal{M}_{L,\mathrm{Q}_1,x}(0^+) \cap \mathcal{M}_{L,\mathrm{Q}_2,x}(0^+) = \emptyset\,.$$

i). Show that L satisfies the assumptions of Corollary 6.8.
ii). Show that for margin-based and distance-based convex losses $L \neq 0$ there exist two distributions Q_1 and Q_2 on Y having mutually distinct L-risk minimizers.

Hint: For *i)* show that there exists a constant $c > 0$ such that for all $x \in X$ we have $\operatorname{dist}(t, \mathcal{M}_{L,\mathrm{Q}_2,x}(0^+)) \geq c$ if $t \in \mathcal{M}_{1,x}$ and $\operatorname{dist}(t, \mathcal{M}_{L,\mathrm{Q}_1,x}(0^+)) \geq c$ if $t \in \mathcal{M}_{2,x}$. Then repeat the argument used in the proof of Lemma 3.15.

6.6. Simple analysis of approximate empirical risk minimizers ($\star\star\star$)
Let $L : X \times Y \times \mathbb{R} \to [0, \infty)$ be a loss function, $B > 0$ be a real number, and $\mathcal{F} \subset \mathcal{L}_0(X)$ be a finite set of bounded measurable functions such that $L(x, y, f(x)) \leq B$ for all $(x, y) \in X \times Y$ and all $f \in \mathcal{F}$. In addition, assume that for some $\epsilon > 0$ we have a measurable learning algorithm that produces *ϵ-approximate* minimizers f_D of $\mathcal{R}_{L,\mathrm{D}}(\,\cdot\,)$, i.e.,

$$\mathcal{R}_{L,\mathrm{D}}(f_D) \leq \inf_{f \in \mathcal{F}} \mathcal{R}_{L,\mathrm{D}}(f) + \epsilon\,, \qquad D \in (X \times Y)^n.$$

Show that, for all $\tau > 0$ and all $n \geq 1$, the following inequality holds:

$$\mathrm{P}^n\biggl(D \in (X\times Y)^n : \mathcal{R}_{L,\mathrm{P}}(f_D) < \mathcal{R}^*_{L,\mathrm{P},\mathcal{F}} + B\sqrt{\frac{2\tau + 2\ln(2|\mathcal{F}|)}{n}} + \epsilon\biggr) \geq 1 - e^{-\tau}\,.$$

6.7. A simple example of overfitting ERM ($\star\star\star$)
Let $(X, \mathcal{A})$ be a measurable space such that $\{x\} \in \mathcal{A}$ for all $x \in X$. Furthermore, let $Y := \{-1, 1\}$, L_{class} be the binary classification loss, and $\mathcal{F} := \mathcal{L}_\infty(X)$. For $D := ((x_1, y_1), \dots, (x_n, y_n)) \in (X\times Y)^n$, define the function $f_D : X \to \mathbb{R}$ by

$$f_D := \frac{1}{n}\sum_{i=1}^n y_i \mathbf{1}_{\{x_i\}}\,.$$

i). Show that $D \mapsto f_D$ is a measurable empirical risk minimizer with respect to $\mathcal{F}$ and L_{class}.
ii). Let P be a distribution on $X \times Y$ such that $\mathrm{P}_X(\{x\}) = 0$ for all $x \in X$. Show that $\mathcal{R}_{L,\mathrm{P}}(f_D) = \mathcal{R}_{L,\mathrm{P}}(0)$.
iii). Find distributions P on $X \times Y$ such that $\mathcal{R}^*_{L,\mathrm{P}} = 0$ and $\mathcal{R}_{L,\mathrm{P}}(f_D) = 1/2$.

6.8. Entropy vs. covering numbers ($\star\star\star$)
Let (T, d) be a metric space and $a > 0$ and $q > 0$ be constants such that

$$\ln \mathcal{N}(T, d, \varepsilon) < \Bigl(\frac{a}{\varepsilon}\Bigr)^q, \qquad \varepsilon > 0.$$

Show that $e_n(T, d) \leq 3^{\frac{1}{q}} a\, n^{-\frac{1}{q}}$ for all $n \geq 1$.

6.9. Consistency and rates for SVMs using their stability ($\star\star$)
Let $L : X \times Y \times \mathbb{R} \to [0, \infty)$ be a convex, Lipschitz continuous loss satisfying $L(x, y, 0) \leq 1$ for all $(x, y) \in X \times Y$, and $|L|_1 \leq 1$. Moreover, let H be a separable RKHS with measurable kernel k over X satisfying $\|k\|_\infty \leq 1$, and let P be a distribution on $X \times Y$ such that H is dense in $L_1(\mathrm{P}_X)$.
i). Show that with probability P^n not less than $1 - e^{-\tau}$ we have

$$\lambda\|f_{\mathrm{D},\lambda}\|_H^2 + \mathcal{R}_{L,\mathrm{P}}(f_{\mathrm{D},\lambda}) - \mathcal{R}^*_{L,\mathrm{P}} < A_2(\lambda) + \lambda^{-1}\biggl(\sqrt{\frac{8\tau}{n}} + \sqrt{\frac{4}{n}} + \frac{8\tau}{3n}\biggr).$$

ii). Show that the SVM is consistent whenever we choose a sequence $(\lambda_n) \subset (0, 1]$ such that $\lim_{n\to\infty} \lambda_n = 0$ and $\lim_{n\to\infty} \lambda_n^2 n = \infty$.
iii). Assume that (6.23) holds, i.e., there exist constants $c > 0$ and $\beta \in (0, 1]$ such that $A_2(\lambda) \leq c\lambda^\beta$ for all $\lambda > 0$. Show that the asymptotically best choice for λ_n is a sequence that behaves like $n^{-\frac{1}{2\beta+2}}$ and that the resulting learning rate is given by

$$\mathrm{P}^n\Bigl(D \in (X \times Y)^n : \mathcal{R}_{L,\mathrm{P}}(f_{\mathrm{D},\lambda_n}) - \mathcal{R}^*_{L,\mathrm{P}} \leq \tilde{C}\tau n^{-\frac{\beta}{2\beta+2}}\Bigr) \geq 1 - e^{-\tau}\,,$$

where $\tilde{C}$ is a constant independent of τ and n.
iv). Show that the learning rates established in *iii)* are faster than those of Theorem 6.25 if $p > 1/(2\beta + 1)$.