7

Advanced Statistical Analysis of SVMs (*)

Overview. *In the previous chapter, we established both consistency and learning rates using relatively simple oracle inequalities. The first goal of this chapter is to illustrate why these learning rates are too loose. We then establish refined oracle inequalities that lead to learning rates that are substantially faster than those of the previous chapter.*

Prerequisites. *The first two sections require only the statistical analysis from Chapter 6. The following two sections additionally need aspects of empirical process theory provided in Sections A.8 and A.9. The final section requires knowledge from Sections A.5.2 and A.5.6.*

Usage. *We will use the derived oracle inequalities in Chapters 8 and 9 when dealing with classification and regression, respectively.*

In the previous chapter, we presented two techniques to establish oracle inequalities for SVMs. We further showed how these oracle inequalities can be used to establish learning rates if assumptions on the approximation error function are made. The first goal of this chapter is to demonstrate in Section 7.1 that these learning rates are almost always suboptimal. We will then present a new technique to establish sharper oracle inequalities. We begin by considering ERM over finite sets of functions since this learning method is, as in Chapter 6, a suitable raw model for studying the basic principles of this technique. The resulting oracle inequality, which will later be used for parameter selection, is presented in Section 7.2. Unlike in the previous chapter, however, there is no simple yet effective way to extend this technique to infinite sets of functions. This forces us to introduce some heavy machinery from empirical process theory, which is summarized in Section A.8. We also need a new concentration inequality, known as Talagrand's inequality, which, unlike the concentration inequalities of the previous chapter, deals with suprema of functions directly. Since the highly non-trivial proof of Talagrand's inequality is out of the scope of this chapter, it is deferred to Section A.9. In Section 7.3, we will carefully introduce these tools in the process of establishing an oracle inequality for ERM over *infinite* sets of functions, so that the reader immediately gets an idea, of how the different tools work together. We then adapt this approach to SVMs and related modifications of ERM in Section 7.4. Finally, we revisit entropy numbers for RKHSs in Section 7.5.

7.1 Why Do We Need a Refined Analysis?

In this section, we show that the oracle inequalities established in the previous chapter almost always lead to suboptimal learning rates. This will give us the motivation to look for more advanced techniques in the following sections.

Let us begin by recalling the situation of Theorem 6.25. To this end, we assume for simplicity that the loss L is Lipschitz continuous with $|L|_1 \leq 1$. In addition, we again assume that there are constants $a \geq 1$ and $p > 0$ such that

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

Theorem 6.25 then implied the oracle inequality (6.21), i.e., for fixed $n \geq 1$, $\lambda \in (0,1]$, and $\tau \geq 1$, we have[1]

$$\lambda \|f_{\mathrm{D},\lambda}\|_H^2 + \mathcal{R}_{L,\mathrm{P}}(f_{\mathrm{D},\lambda}) - \mathcal{R}^*_{L,\mathrm{P}} < A_2(\lambda) + \frac{3}{\lambda^{1/2}} \left(2\Big(\frac{2a}{n}\Big)^{\frac{1}{2+2p}} + \Big(\frac{2\tau}{n}\Big)^{\frac{1}{2}} \right) \tag{7.1}$$

with probability P^n not less than $1 - e^{-\tau}$. The positive aspect of this oracle inequality is that it holds for *all* distributions P on $X \times Y$. From the machine learning perspective, this distribution independence is highly desirable since one of its basic assumptions is that the data-generating distribution P is unknown. However, this distribution independence is also the weakness of the oracle inequality above since for most distributions it is overly pessimistic in the sense that the resulting learning rates are suboptimal. To explain this, let us assume for simplicity that for sample size n we have chosen a regularization parameter $\lambda_n \in (0,1]$. Now recall that in the proof of Theorem 6.25 we used the trivial estimate

$$\|f_{\mathrm{D},\lambda_n}\|_H \leq \lambda_n^{-1/2} \tag{7.2}$$

to determine the function class over which the SVM actually minimizes. However, (7.1) then shows that with high probability we have $\|f_{\mathrm{D},\lambda_n}\|_H \leq (\varepsilon_n/\lambda_n)^{1/2}$, where ε_n is a shorthand for the right-hand side of (7.1). Assuming that we have chosen λ_n such that $\varepsilon_n \to 0$ for $n \to \infty$, we hence see that with high probability we have an estimate on $\|f_{\mathrm{D},\lambda_n}\|_H$ that is sharper than (7.2) for large n. To refine the analysis of Theorem 6.25, we could now exclude in the proof of Theorem 6.25 the set of samples D where this sharper estimate is not satisfied. As a consequence, we would work with a smaller function class and with a smaller bound B on the suprema. Now recall that both the size of the function class and B have a significant impact on the oracle inequality of Theorem 6.25 and hence on (7.1). To illustrate this, assume that we have chosen $\lambda_n := n^{-\gamma}$ for some $0 < \gamma < 1/(1+p)$ and all $n \geq 1$. Moreover,

[1] Here, as in the rest of this chapter, we assume that $(X \times Y)^n$ is equipped with the universal completion of the product σ-algebra of $(X \times Y)^n$. In addition, P^n denotes the canonical extension of the n-fold product measure of P to this completion. Recall that these conventions together with Lemmas 6.23 and 6.3 make it possible to ignore measurability questions for SVMs.

assume that $A_2(\lambda) \leq c\lambda^\beta$ for some constants $c > 0$, $\beta \in (0, 1]$, and all $\lambda > 0$. By following the path above, we would then obtain an improvement of (7.1) by a factor of the form $n^{-\alpha}$ for some $\alpha > 0$. This discussion shows that the original estimate (7.1) indeed leads to suboptimal learning rates. Moreover, it shows another dilemma: the improved version of (7.1) leads directly to a further improvement of (7.2), which in turn yields a further improvement of (7.1), and so on. On the other hand, it is not hard to see that such an iteration would increase the arising constants since we have to exclude more and more sets of small probability, and hence it seems likely that an analysis following this path would be rather technical. In addition, it would require choosing λ_n *a priori*, and hence we would not obtain an oracle inequality that can be used for the analysis of parameter selection procedures such as the TV-SVM.

Interestingly, the phenomenon above is not the only source for the general suboptimality of Theorem 6.25. However, the second source is more involved, and therefore we only illustrate it for ERM.[2] To this end, let us fix a finite set $\mathcal{F} \subset \mathcal{L}_\infty(X)$ of functions and a loss function $L : X \times Y \times \mathbb{R} \to [0, \infty)$ such that $L(x, y, f(x)) \leq B$ for some constant $B > 0$ and all $(x, y) \in X \times Y$ and $f \in \mathcal{F}$. In addition, we assume that P is a distribution for which there exists an $f^* \in \mathcal{F}$ with $\mathcal{R}_{L,\mathrm{P}}(f^*) = 0$. This implies $L(x, y, f^*(x)) = 0$ for P-almost all $(x, y) \in X \times Y$, and hence we have $\mathcal{R}_{L,\mathrm{D}}(f^*) = 0$ for P^n-almost all $D \in (X \times Y)^n$. From this we conclude that $\mathcal{R}_{L,\mathrm{D}}(f_D) = 0$ almost surely. For $f \in \mathcal{F}$, we now define $h_f \in \mathcal{L}_\infty(X \times Y)$ by

$$h_f(x, y) := L(x, y, f(x)) \,, \qquad (x, y) \in X \times Y \,.$$

Since L is non-negative, this definition immediately yields

$$\mathbb{E}_\mathrm{P} h_f^2 \;\leq\; B\mathbb{E}_\mathrm{P} h_f \,. \tag{7.3}$$

Furthermore, for $r > 0$ and $f \in \mathcal{F}$, we define

$$g_{f,r} := \frac{\mathbb{E}_\mathrm{P} h_f - h_f}{\mathbb{E}_\mathrm{P} h_f + r} \,.$$

For $f \in \mathcal{F}$ with $\mathbb{E}_\mathrm{P} h_f = 0$, we then have $\mathbb{E}_\mathrm{P} h_f^2 = 0$ by (7.3) and hence we obtain $\mathbb{E}_\mathrm{P} g_{f,r}^2 = 0 \leq \frac{B}{2r}$. Moreover, for $f \in \mathcal{F}$ with $\mathbb{E}_\mathrm{P} h_f \neq 0$, we find

$$\mathbb{E}_\mathrm{P} g_{f,r}^2 \leq \frac{\mathbb{E}_\mathrm{P} h_f^2}{(\mathbb{E}_\mathrm{P} h_f + r)^2} \leq \frac{\mathbb{E}_\mathrm{P} h_f^2}{2r\mathbb{E}_\mathrm{P} h_f} \leq \frac{B}{2r} \,,$$

where we used (7.3) and the trivial estimate $2ab \leq (a + b)^2$ for $a, b \geq 0$. In addition, we have

$$\|g_{f,r}\|_\infty = \sup_{(x,y)\in X\times Y} \left| \frac{\mathbb{E}_\mathrm{P} h_f - h_f(x, y)}{\mathbb{E}_\mathrm{P} h_f + r} \right| = \frac{\|\mathbb{E}_\mathrm{P} h_f - h_f\|_\infty}{\mathbb{E}_\mathrm{P} h_f + r} \leq \frac{B}{r}$$

[2] At the end of Section 7.4, we see that this phenomenon indeed occurs for SVMs.

and $\mathrm{E}_{\mathrm{P}} g_{f,r} = 0$. Applying Bernstein's inequality and the union bound, we hence obtain

$$\mathrm{P}^n \left(D \in (X \times Y)^n : \sup_{f \in \mathcal{F}} \mathbb{E}_{\mathrm{D}} g_{f,r} \geq \sqrt{\frac{B\tau}{nr}} + \frac{2B\tau}{3nr} \right) \leq |\mathcal{F}| e^{-\tau} \,,$$

and since $f_D \in \mathcal{F}$, the definition of $g_{f_D,r}$ thus yields

$$\mathrm{P}^n \left(D \in (X \times Y)^n : \mathbb{E}_{\mathrm{P}} h_{f_D} - \mathbb{E}_{\mathrm{D}} h_{f_D} \geq \left(\mathbb{E}_{\mathrm{P}} h_{f_D} + r \right) \left(\sqrt{\frac{B\tau}{nr}} + \frac{2B\tau}{3nr} \right) \right) \leq |\mathcal{F}| e^{-\tau}.$$

Because $\mathbb{E}_{\mathrm{D}} h_{f_D} = \mathcal{R}_{L,\mathrm{D}}(f_D) = 0$ almost surely, we hence conclude that

$$\mathrm{P}^n \left(D \in (X \times Y)^n : \left(1 - \sqrt{\frac{B\tau}{nr}} - \frac{2B\tau}{3nr} \right) \mathbb{E}_{\mathrm{P}} h_{f_D} \geq \sqrt{\frac{rB\tau}{n}} + \frac{2B\tau}{3n} \right) \leq |\mathcal{F}| e^{-\tau}.$$

For $r := \frac{4B\tau}{n}$, the relation $\mathbb{E}_{\mathrm{P}} h_{f_D} = \mathcal{R}_{L,\mathrm{P}}(f_D)$ thus yields

$$\mathrm{P}^n \left(D \in (X \times Y)^n : \mathcal{R}_{L,\mathrm{P}}(f_D) \geq \frac{8B\tau}{n} \right) \leq |\mathcal{F}| e^{-\tau}. \tag{7.4}$$

Compared with (6.15), the latter estimate replaces the square root term $B(\frac{2\tau}{n})^{1/2}$ by the substantially faster decaying linear term $\frac{8B\tau}{n}$. In other words, for the specific type of distribution considered above, our analysis of ERM in Chapter 6 is too loose by a factor of $n^{-1/2}$.

The reason for this improvement is the *variance bound* (7.3), which guarantees a small variance whenever we have a small error $\mathcal{R}_{L,\mathrm{P}}(f) = \mathbb{E}_{\mathrm{P}} h_f$. Interestingly, such a variance bound can also easily be used in a brute-force analysis that does not start with an initial small error. Indeed, let us assume that we would begin our analysis by using Bernstein's inequality together with the trivial variance bound $\mathbb{E}_{\mathrm{P}} h_{f_D}^2 \leq B^2$. We would then obtain a small upper bound on $\mathbb{E}_{\mathrm{P}} h_{f_D}$ that holds with high probability. Using (7.3), this would give us a smaller bound on $\mathbb{E}_{\mathrm{P}} h_{f_D}^2$, which in turn would improve our first bound on $\mathbb{E}_{\mathrm{P}} h_{f_D}$ by another application of Bernstein's inequality. Obviously, this brute-force analysis would be very similar to the iterative proof procedure we discussed around (7.2). This observation suggests that Theorem 6.25 is also suboptimal if P guarantees a variance bound in the sense of (7.3), and we will see in Section 7.4 that this is indeed the case. Remarkably, however, the argument that led to (7.4) avoided this iterative argument by considering the function $g_{f,r}$ in Bernstein's inequality. This trick, in a refined form, will be *the* central idea for the refined analysis of this chapter.

7.2 A Refined Oracle Inequality for ERM

The goal of this section is to generalize the technique of using a variance bound in conjunction with Bernstein's inequality to derive oracle inequalities

for ERM. Although these generalizations still will not be powerful enough to deal with SVMs, they will already illustrate the key ideas. In addition, the derived oracle inequality for ERM will later be used for investigating the adaptivity of data-dependent parameter selection strategies such as the one of TV-SVMs.

Let us begin with the following elementary and widely known lemma.

Lemma 7.1. *For $q \in (1,\infty)$, define $q' \in (1,\infty)$ by $\frac{1}{q} + \frac{1}{q'} = 1$. Then we have*

$$ab \le \frac{a^q}{q} + \frac{b^{q'}}{q'}$$

and $(qa)^{2/q}(q'b)^{2/q'} \le (a+b)^2$ for all $a, b \ge 0$.

Proof. For $a = 0$, the first assertion is trivial, and hence it suffices to consider the case $a > 0$. We define $h_a(b) := a^q/q + b^{q'}/q' - ab$, $b \ge 0$. Obviously, the derivative of this function is $h_a'(b) = b^{q'-1} - a$, and hence h_a has a unique global minimum at $b^* := a^{1/(q'-1)}$. Using $q'/(q'-1) = q$, we find $h_a(b^*) = 0$, which then gives the desired first inequality. The second inequality then follows from the first by a simple variable transformation and $a^2 + b^2 \le (a+b)^2$. □

Before we present the improved oracle inequality for ERM, let us introduce the shorthand $L \circ f$ for the function $(x,y) \mapsto L(x,y,f(x))$, where $f : X \to \mathbb{R}$ is an arbitrary function and $L : X \times Y \times \mathbb{R} \to [0,\infty)$ is a loss.

Theorem 7.2 (Improved oracle inequality for ERM). *Consider a measurable ERM with respect to the loss $L : X \times Y \times \mathbb{R} \to [0,\infty)$ and the finite set $\mathcal{F} \subset \mathcal{L}_0(X)$. Moreover, let P be a distribution on $X \times Y$ that has a Bayes decision function $f^*_{L,\mathrm{P}}$. Assume that there exist constants $B > 0$, $\vartheta \in [0,1]$, and $V \ge B^{2-\vartheta}$ such that for all $f \in \mathcal{F}$ we have*

$$\|L \circ f - L \circ f^*_{L,\mathrm{P}}\|_\infty \le B\,, \tag{7.5}$$

$$\mathbb{E}_\mathrm{P}\big(L \circ f - L \circ f^*_{L,\mathrm{P}}\big)^2 \le V \cdot \big(\mathbb{E}_\mathrm{P}(L \circ f - L \circ f^*_{L,\mathrm{P}})\big)^\vartheta\,. \tag{7.6}$$

Then, for all fixed $\tau > 0$ and $n \ge 1$, we have with probability P^n not less than $1 - e^{-\tau}$ that

$$\mathcal{R}_{L,\mathrm{P}}(f_D) - \mathcal{R}^*_{L,\mathrm{P}} \;<\; 6\big(\mathcal{R}^*_{L,\mathrm{P},\mathcal{F}} - \mathcal{R}^*_{L,\mathrm{P}}\big) + 4\left(\frac{8V\big(\tau + \ln(1+|\mathcal{F}|)\big)}{n}\right)^{\frac{1}{2-\vartheta}}.$$

Proof. We first note that since $\mathcal{R}_{L,\mathrm{P}}(f_D) - \mathcal{R}^*_{L,\mathrm{P}} \le B$ and $V \ge B^{2-\vartheta}$, it suffices to consider the case $n \ge 8\tau$. For $f \in \mathcal{F}$ we define $h_f := L \circ f - L \circ f^*_{L,\mathrm{P}}$, and in addition we fix an $f_0 \in \mathcal{F}$. Since $\mathcal{R}_{L,\mathrm{D}}(f_D) \le \mathcal{R}_{L,\mathrm{D}}(f_0)$, we then have $\mathbb{E}_\mathrm{D} h_{f_D} \le \mathbb{E}_\mathrm{D} h_{f_0}$, and consequently we obtain

$$\begin{aligned}\mathcal{R}_{L,\mathrm{P}}(f_D) - \mathcal{R}_{L,\mathrm{P}}(f_0) &= \mathbb{E}_\mathrm{P} h_{f_D} - \mathbb{E}_\mathrm{P} h_{f_0} \\ &\le \mathbb{E}_\mathrm{P} h_{f_D} - \mathbb{E}_\mathrm{D} h_{f_D} + \mathbb{E}_\mathrm{D} h_{f_0} - \mathbb{E}_\mathrm{P} h_{f_0} \end{aligned} \tag{7.7}$$

for all $D \in (X \times Y)^n$. Let us first estimate $\mathbb{E}_D h_{f_0} - \mathbb{E}_P h_{f_0}$ for the case $\vartheta > 0$. To this end, we observe that $\|h_{f_0} - \mathbb{E}_P h_{f_0}\|_\infty \leq 2B$ and

$$\mathbb{E}_P(h_{f_0} - \mathbb{E}_P h_{f_0})^2 \leq \mathbb{E}_P h_{f_0}^2 \leq V(\mathbb{E}_P h_{f_0})^\vartheta .$$

In addition, for $q := \frac{2}{2-\vartheta}$, $q' := \frac{2}{\vartheta}$, $a := \left(\frac{2^{1-\vartheta}\vartheta^\vartheta V\tau}{n}\right)^{1/2}$, and $b := \left(\frac{2\mathbb{E}_P h_{f_0}}{\vartheta}\right)^{\vartheta/2}$, Lemma 7.1 shows that

$$\sqrt{\frac{2\tau V(\mathbb{E}_P h_{f_0})^\vartheta}{n}} \leq \left(1-\frac{\vartheta}{2}\right)\left(\frac{2^{1-\vartheta}\vartheta^\vartheta V\tau}{n}\right)^{\frac{1}{2-\vartheta}} + \mathbb{E}_P h_{f_0} \leq \left(\frac{2V\tau}{n}\right)^{\frac{1}{2-\vartheta}} + \mathbb{E}_P h_{f_0},$$

and hence Bernstein's inequality shows that we have

$$\mathbb{E}_D h_{f_0} - \mathbb{E}_P h_{f_0} < \mathbb{E}_P h_{f_0} + \left(\frac{2V\tau}{n}\right)^{\frac{1}{2-\vartheta}} + \frac{4B\tau}{3n} \tag{7.8}$$

with probability P^n not less than $1 - e^{-\tau}$. Furthermore, note that, for $\vartheta = 0$, the same inequality holds by Hoeffding's inequality and $\|h_{f_0}\|_\infty \leq B \leq \sqrt{V}$.

To estimate the remaining term $\mathbb{E}_P h_{f_D} - \mathbb{E}_D h_{f_D}$, we define the functions

$$g_{f,r} := \frac{\mathbb{E}_P h_f - h_f}{\mathbb{E}_P h_f + r}, \qquad f \in \mathcal{F},\, r > 0 .$$

Obviously, we have $\|g_{f,r}\|_\infty \leq 2Br^{-1}$. Moreover, for $\vartheta > 0$, $b := \mathbb{E}_P h_f \neq 0$, $q := \frac{2}{2-\vartheta}$, $q' := \frac{2}{\vartheta}$, and $a := r$, the second inequality of Lemma 7.1 yields

$$\mathbb{E}_P g_{f,r}^2 \leq \frac{\mathbb{E}_P h_f^2}{(\mathbb{E}_P h_f + r)^2} \leq \frac{(2-\vartheta)^{2-\vartheta}\vartheta^\vartheta\, \mathbb{E}_P h_f^2}{4r^{2-\vartheta}(\mathbb{E}_P h_f)^\vartheta} \leq V r^{\vartheta-2} .$$

Furthermore, for $\vartheta > 0$ and $\mathbb{E}_P h_f = 0$, we have $\mathbb{E}_P h_f^2 = 0$ by (7.6), which in turn implies $\mathbb{E}_P g_{f,r}^2 \leq V r^{\vartheta-2}$. Finally, in the case $\vartheta = 0$, we easily obtain $\mathbb{E}_P g_{f,r}^2 \leq \mathbb{E}_P h_f^2\, r^{-2} \leq V r^{\vartheta-2}$, and hence we have $\mathbb{E}_P g_{f,r}^2 \leq V r^{\vartheta-2}$ in all cases. Consequently, Bernstein's inequality yields

$$P^n\left(D \in (X \times Y)^n : \sup_{f\in\mathcal{F}} \mathbb{E}_D g_{f,r} < \sqrt{\frac{2V\tau}{nr^{2-\vartheta}}} + \frac{4B\tau}{3nr}\right) \;\geq\; 1 - |\mathcal{F}|e^{-\tau} \tag{7.9}$$

for all $r > 0$. Now observe that for $D \in (X \times Y)^n$ satisfying

$$\sup_{f\in\mathcal{F}} \mathbb{E}_D g_{f,r} < \sqrt{\frac{2V\tau}{nr^{2-\vartheta}}} + \frac{4B\tau}{3nr},$$

the definition of $g_{f_D,r}$ and $f_D \in \mathcal{F}$ imply

$$\mathbb{E}_P h_{f_D} - \mathbb{E}_D h_{f_D} < \mathbb{E}_P h_{f_D}\left(\sqrt{\frac{2V\tau}{nr^{2-\vartheta}}} + \frac{4B\tau}{3nr}\right) + \sqrt{\frac{2V\tau r^\vartheta}{n}} + \frac{4B\tau}{3n} .$$

By combining this estimate with (7.7), (7.8), and (7.9), we thus see that

$$\mathbb{E}_{\mathrm{P}} h_{f_D} < 2\mathbb{E}_{\mathrm{P}} h_{f_0} + \mathbb{E}_{\mathrm{P}} h_{f_D} \left(\sqrt{\frac{2V\tau}{nr^{2-\vartheta}}} + \frac{4B\tau}{3nr} \right) + \sqrt{\frac{2V\tau r^{\vartheta}}{n}} + \left(\frac{2V\tau}{n} \right)^{\frac{1}{2-\vartheta}} + \frac{8B\tau}{3n}$$

holds with probability P^n not less than $1 - (1 + |\mathcal{F}|)e^{-\tau}$. Let us now define $r := \left(\frac{8V\tau}{n}\right)^{1/(2-\vartheta)}$. Then we obviously have

$$\sqrt{\frac{2V\tau}{nr^{2-\vartheta}}} = \frac{1}{2} \qquad \text{and} \qquad \sqrt{\frac{2V\tau r^{\vartheta}}{n}} = \frac{r}{2}\,,$$

and by using $V \geq B^{2-\vartheta}$ and $n \geq 8\tau$ we also find

$$\frac{4B\tau}{3nr} \;=\; \frac{1}{6} \cdot \frac{8\tau}{n} \cdot \frac{B}{r} \;\leq\; \frac{1}{6} \cdot \left(\frac{8\tau}{n}\right)^{\frac{1}{2-\vartheta}} \cdot \frac{V^{\frac{1}{2-\vartheta}}}{r} \;=\; \frac{1}{6}$$

and $\frac{8B\tau}{3n} \leq \frac{r}{3}$. In addition, $2 \leq 4^{\frac{1}{2-\vartheta}}$ and the definition of r yield $(\frac{2V\tau}{n})^{\frac{1}{2-\vartheta}} \leq \frac{r}{2}$, and hence we have with probability P^n not less than $1 - (1 + |\mathcal{F}|)e^{-\tau}$ that

$$\mathbb{E}_{\mathrm{P}} h_{f_D} < 2\mathbb{E}_{\mathrm{P}} h_{f_0} + \frac{2}{3}\mathbb{E}_{\mathrm{P}} h_{f_D} + \frac{4}{3}r\,.$$

We now obtain the assertion by some simple algebraic transformations and taking a function $f_0 \in \mathcal{F}$ such that $\mathcal{R}_{L,\mathrm{P}}(f_0) = \min\{\mathcal{R}_{L,\mathrm{P}}(f) : f \in \mathcal{F}\}$. □

Note that in the proof of Theorem 7.2 we did not strive to obtain the smallest possible constants. Instead we tried to keep both the proof and the oracle inequality as simple as possible.

Obviously, the crucial assumption of Theorem 7.2 is the variance bound (7.6). Unfortunately, for many loss functions it is a non-trivial task to establish such a bound, as we will see in Sections 8.3 and 9.5, where we consider this issue for the hinge loss and the pinball loss, respectively. On the other hand, the following example shows that for the least squares loss and *bounded* Y, a variance bound always holds.

Example 7.3. Let $M > 0$ and $Y \subset [-M, M]$ be a closed subset. Moreover, let L be the **least squares loss**, X be a non-empty set equipped with some σ-algebra, and P be a distribution on $X \times Y$. By the non-negativity of L and $f^*_{L,\mathrm{P}}(x) = \mathbb{E}_{\mathrm{P}}(Y|x) \in [-M, M]$, $x \in X$, we first observe that

$$\left|L(y, f(x)) - L(y, f^*_{L,\mathrm{P}}(x))\right| \leq \sup_{y', t \in [-M,M]} (y' - t)^2 = 4M^2$$

for all measurable $f : X \to [-M, M]$ and all $(x, y) \in X \times Y$. Consequently, (7.5) holds for $B := 4M^2$. Moreover, we also have

$$\begin{aligned}\left(L(y, f(x)) - L(y, f^*_{L,\mathrm{P}}(x))\right)^2 &= \left((f(x) + f^*_{L,\mathrm{P}}(x) - 2y)(f(x) - f^*_{L,\mathrm{P}}(x))\right)^2 \\ &\leq 16M^2 \left(f(x) - f^*_{L,\mathrm{P}}(x)\right)^2,\end{aligned}$$

and hence we find

$$\mathbb{E}_{\mathrm{P}}(L \circ f - L \circ f^*_{L,\mathrm{P}})^2 \leq 16M^2\, \mathbb{E}_{\mathrm{P}}(f - f^*_{L,\mathrm{P}})^2 \tag{7.10}$$

$$= 16M^2\, \mathbb{E}_{\mathrm{P}}(L \circ f - L \circ f^*_{L,\mathrm{P}}) \tag{7.11}$$

In other words, (7.6) holds for $V := 16M^2$ and $\vartheta = 1$. ◁

Note that the variance bound established in the preceding example holds for the optimal exponent $\vartheta = 1$, which leads to a $1/n$ behavior of the oracle inequality presented in Theorem 7.2. Besides this, however, the preceding example also provides us with a "template approach" for establishing variance bounds. Indeed, (7.10) only uses the local Lipschitz continuity of the least squares loss when restricted to label domain Y, while (7.11) is a very special case of self-calibration. Interestingly, all later established variance bounds will essentially follow this pattern of combining Lipschitz continuity with self-calibration. Unlike for the least squares loss, which is nicely self-calibrated for *all* distributions having bounded label space Y, for most other interesting losses, establishing non-trivial self-calibration properties requires identifying suitable distributions. Unfortunately, in many cases this is a non-trivial task.

7.3 Some Advanced Machinery

One of the main ideas in the proof of Theorem 7.2 was to apply Bernstein's inequality to functions of the form

$$g_{f,r} := \frac{\mathbb{E}_{\mathrm{P}} h_f - h_f}{\mathbb{E}_{\mathrm{P}} h_f + r}\,, \qquad f \in \mathcal{F},\, r > 0\,, \tag{7.12}$$

where $h_f := L \circ f - L \circ f^*_{L,\mathrm{P}}$ and $\mathcal{F}$ was a *finite* set of functions. However, we have already seen in Chapter 6 that the statistical analysis of SVMs requires *infinite* sets of functions, namely balls of the RKHS used. Now a straightforward generalization of Theorem 7.2 to such $\mathcal{F}$ would assume $\mathcal{N}(\mathcal{F}, \|\cdot\|_\infty, \varepsilon) < \infty$ for all $\varepsilon > 0$ (see Exercise 7.3 for the details of such an extension). However, we will see later in Section 7.4 (see also Exercise 7.5) that the covering numbers with respect to the supremum norm often are too large. This suboptimality motivates this section, whose goal is to present more sophisticated tools for generalizing Theorem 7.2 to infinite function classes. However, these tools require much more involved proofs than the ones seen so far, and hence some of the more complicated results and their proofs are presented in Sections A.8 and A.9. By postponing these parts, we hope to give the reader a better understanding of how these tools work together.

Since most of the following results involve expectations of suprema over *uncountable* sets, we first clarify measurability issues. To this end, we introduce the following notion.

Definition 7.4. *Let (T,d) be a metric space and $(Z,\mathcal{A})$ be a measurable space. A family of maps $(g_t)_{t\in T}\subset\mathcal{L}_0(Z)$ is called a* ***Carathéodory family*** *if $t\mapsto g_t(z)$ is continuous for all $z\in Z$. Moreover, if T is separable or complete, we say that $(g_t)_{t\in T}$ is* ***separable*** *or* ***complete****, respectively.*

In the following we call a subset $\mathcal{G}\subset\mathcal{L}_0(Z)$ a **(separable or complete) Carathéodory set** if there exists a (separable or complete) metric space (T,d) and a Carathéodory family $(g_t)_{t\in T}\subset\mathcal{L}_0(Z)$ such that $\mathcal{G}=\{g_t : t\in T\}$. Note that, by the continuity of $t\mapsto g_t(z)$, Carathéodory sets satisfy

$$\sup_{g\in\mathcal{G}} g(z) = \sup_{t\in T} g_t(z) = \sup_{t\in S} g_t(z)\,, \qquad z\in Z, \tag{7.13}$$

for all dense $S\subset T$. In particular, for separable Carathéodory sets $\mathcal{G}$, there exists a *countable* and dense $S\subset T$, and hence the map $z\mapsto\sup_{t\in T} g_t(z)$ is measurable for such $\mathcal{G}$. Finally, recall Lemma A.3.17, which shows that the map $(z,t)\mapsto g_t(z)$ is measurable if T is separable *and* complete.

Now our first result generalizes Bernstein's inequality to suprema of separable Carathéodory sets.

Theorem 7.5 (Simplified Talagrand's inequality). *Let $(Z,\mathcal{A},\mathrm{P})$ be a probability space and $\mathcal{G}\subset\mathcal{L}_0(Z)$ be a separable Carathéodory set. Furthermore, let $B\geq 0$ and $\sigma\geq 0$ be constants such that $\mathbb{E}_{\mathrm{P}}g=0$, $\mathbb{E}_{\mathrm{P}}g^2\leq\sigma^2$, and $\|g\|_\infty\leq B$ for all $g\in\mathcal{G}$. For $n\geq 1$, we define $G:Z^n\to\mathbb{R}$ by*

$$G(z_1,\ldots,z_n) := \sup_{g\in\mathcal{G}}\left|\frac{1}{n}\sum_{j=1}^n g(z_j)\right|, \qquad z=(z_1,\ldots,z_n)\in Z^n.$$

Then, for all $\tau>0$ and all $\gamma>0$, we have

$$\mathrm{P}^n\left(\left\{z\in Z^n : G(z)\geq(1+\gamma)\mathbb{E}_{\mathrm{P}^n}G+\sqrt{\frac{2\tau\sigma^2}{n}}+\left(\frac{2}{3}+\frac{1}{\gamma}\right)\frac{\tau B}{n}\right\}\right)\;\leq\; e^{-\tau}\,.$$

Proof. By (7.13), we may assume without loss of generality that $\mathcal{G}$ is countable. For fixed $a,b>0$ we now define $h:(0,\infty)\to(0,\infty)$ by $h(\gamma):=\gamma a+\gamma^{-1}b$, $\gamma>0$. Then elementary calculus shows that h has a global minimum at $\gamma^*:=\sqrt{b/a}$, and consequently we have $2\sqrt{ab}=h(\gamma^*)\leq h(\gamma)=\gamma a+\gamma^{-1}b$ for all $\gamma>0$. From this we conclude that

$$\sqrt{\frac{2\tau(\sigma^2+2B\mathbb{E}_{\mathrm{P}^n}G)}{n}}\leq\sqrt{\frac{2\tau\sigma^2}{n}}+2\sqrt{\frac{\tau B\mathbb{E}_{\mathrm{P}^n}G}{n}}\leq\sqrt{\frac{2\tau\sigma^2}{n}}+\gamma\mathbb{E}_{\mathrm{P}^n}G+\frac{\tau B}{\gamma n}\,,$$

and hence we obtain the assertion by applying Talagrand's inequality stated in Theorem A.9.1. □

Following the idea of the proof of Theorem 7.2, our first goal is to apply the preceding theorem to the family of maps $(g_{f,r})_{f\in\mathcal{F}}$, where $g_{f,r}$ is defined by (7.12). To this end, we first show that this is a separable Carathéodory family.

Lemma 7.6. *Let $L : X \times Y \times \mathbb{R} \to [0,\infty)$ be a continuous loss, P be a distribution on $X \times Y$, and $f^*_{L,\mathrm{P}}$ be a Bayes decision function. Moreover, let $\mathcal{F} \subset \mathcal{L}_0(X)$ be equipped with a separable metric that dominates the pointwise convergence in the sense of (2.8). If there exists a constant $B > 0$ such that $\|L \circ f - L \circ f^*_{L,\mathrm{P}}\|_\infty \leq B$ for all $f \in \mathcal{F}$, then $(g_{f,r})_{f\in\mathcal{F}}$, where $g_{f,r}$ is defined by (7.12) and $r > 0$ is fixed, is a separable Carathéodory family.*

Proof. Since the metric of $\mathcal{F}$ dominates the pointwise convergence we see that for fixed $(x,y) \in X \times Y$ the $\mathbb{R}$-valued map $f \mapsto f(x)$ defined on $\mathcal{F}$ is continuous. Using the continuity of L, it is then easy to conclude that $(h_f)_{f\in\mathcal{F}}$, where $h_f := L\circ f - L\circ f^*_{L,\mathrm{P}}$ is a Carathéodory family. Moreover, we have $\|h_f\|_\infty \leq B$ for all $f \in \mathcal{F}$, and hence Lebesgue's dominated convergence theorem shows that $f \mapsto \mathbb{E}_\mathrm{P} h_f$ is continuous. From this we obtain the assertion. □

To ensure that $(g_{f,r})_{f\in\mathcal{F}}$ is a separable Carathéodory family, we assume in the rest of this section that the assumptions of Lemma 7.6 are satisfied.

Let us now consider the other assumptions of Theorem 7.5. To this end, recall that we have already seen in the proof of Theorem 7.2 that bounds of the form (7.5) and (7.6) lead to the estimates $\|g_{f,r}\|_\infty \leq Br^{-1}$ and $\mathbb{E}_\mathrm{P} g^2_{f,r} \leq Vr^{\vartheta-2}$, respectively. Moreover, we obviously have $\mathbb{E}_\mathrm{P} g_{f,r} = 0$. Applying Theorem 7.5 for $\gamma = 1$, we hence see that for all $\tau > 0$ we have

$$\sup_{f\in\mathcal{F}} \frac{\mathbb{E}_\mathrm{P} h_f - \mathbb{E}_\mathrm{D} h_f}{\mathbb{E}_\mathrm{P} h_f + r} < 2\mathbb{E}_{D\sim\mathrm{P}^n} \sup_{f\in\mathcal{F}} \frac{\mathbb{E}_\mathrm{P} h_f - \mathbb{E}_\mathrm{D} h_f}{\mathbb{E}_\mathrm{P} h_f + r} + \sqrt{\frac{2V\tau}{nr^{2-\vartheta}}} + \frac{5B\tau}{3nr} \tag{7.14}$$

with probability P^n not less than $1 - e^{-\tau}$. In other words, besides the expectation on the right-hand side, we have basically obtained the key estimate (7.9) of the proof of Theorem 7.2 for general, *not necessarily finite* sets $\mathcal{F}$. The rest of this section is thus devoted to techniques for bounding the additional expectation. We begin with a method that removes the denominator.

Theorem 7.7 (Peeling). *Let $(Z, \mathcal{A}, \mathrm{P})$ be a probability space, (T,d) be a separable metric space, $h : T \to [0,\infty)$ be a continuous function, and $(g_t)_{t\in T} \subset \mathcal{L}_0(Z)$ be a Carathéodory family. We define $r^* := \inf\{h(t) : t \in T\}$. Moreover, let $\varphi : (r^*,\infty) \to [0,\infty)$ be a function such that $\varphi(4r) \leq 2\varphi(r)$ and*

$$\mathbb{E}_{z\sim\mathrm{P}} \sup_{\substack{t\in T\\ h(t)\leq r}} |g_t(z)| \;\leq\; \varphi(r)$$

for all $r > r^*$. Then, for all $r > r^*$, we have

$$\mathbb{E}_{z\sim\mathrm{P}} \sup_{t\in T} \frac{g_t(z)}{h(t) + r} \;\leq\; \frac{4\varphi(r)}{r}\,.$$

Proof. For $z \in Z$ and $r > r^*$, we have

$$\sup_{t\in T} \frac{g_t(z)}{h(t)+r} \leq \sup_{\substack{t\in T\\ h(t)\leq r}} \frac{|g_t(z)|}{r} + \sum_{i=0}^{\infty} \sup_{\substack{t\in T\\ h(t)\in[r4^i, r4^{i+1}]}} \frac{|g_t(z)|}{r4^i + r}\,,$$

where we used the convention $\sup \emptyset := 0$. Consequently, we obtain

$$\mathbb{E}_{z\sim \mathrm{P}} \sup_{t\in T} \frac{g_t(z)}{h(t)+r} \le \frac{\varphi(r)}{r} + \frac{1}{r}\sum_{i=0}^{\infty} \frac{1}{4^i+1}\, \mathbb{E}_{z\sim \mathrm{P}} \sup_{\substack{t\in T \\ h(t)\le r4^{i+1}}} |g_t(z)|$$
$$\le \frac{1}{r}\left(\varphi(r) + \sum_{i=0}^{\infty} \frac{\varphi(r4^{i+1})}{4^i+1}\right).$$

Moreover, induction yields $\varphi(r4^{i+1}) \le 2^{i+1}\varphi(r)$, $i \ge 0$, and hence we obtain

$$\mathbb{E}_{z\sim \mathrm{P}} \sup_{t\in T} \frac{g_t(z)}{h(t)+r} \le \frac{\varphi(r)}{r}\left(1 + \sum_{i=0}^{\infty} \frac{2^{i+1}}{4^i+1}\right) \le \frac{\varphi(r)}{r}\left(1 + \frac{2}{1+1} + \sum_{i=1}^{\infty} 2^{-i+1}\right).$$

From this estimate, we easily obtain the assertion. □

Note that the preceding proof actually yields a constant slightly smaller than 4. Indeed, numerical evaluation suggests a value of approximately 3.77, but since the goal of this section is to present the general picture and not a path that leads to the smallest possible constants, we will work with the slightly larger but simpler constant.

Let us now return to the concentration inequality (7.14) we derived from Talagrand's inequality. To apply the peeling argument, we write

$$\mathcal{H}_r := \left\{h_f : f \in \mathcal{F} \text{ and } \mathbb{E}_{\mathrm{P}} h_f \le r\right\}, \qquad r > 0. \tag{7.15}$$

Furthermore, for $r^* := \inf\{r > 0 : \mathcal{H}_r \neq \emptyset\} = \inf\{\mathbb{E}_{\mathrm{P}} h_f : f \in \mathcal{F}\}$, we assume that we have a function $\varphi_n : (r^*, \infty) \to [0, \infty)$ satisfying $\varphi_n(4r) \le 2\varphi_n(r)$ and

$$\mathbb{E}_{D\sim \mathrm{P}^n} \sup_{h\in\mathcal{H}_r} \left|\mathbb{E}_{\mathrm{P}} h - \mathbb{E}_{\mathrm{D}} h\right| \;\le\; \varphi_n(r) \tag{7.16}$$

for all $r > r^*$. By Theorem 7.7 and (7.14), we then see that we have

$$\sup_{f\in\mathcal{F}} \frac{\mathbb{E}_{\mathrm{P}} h_f - \mathbb{E}_{\mathrm{D}} h_f}{\mathbb{E}_{\mathrm{P}} h_f + r} < \frac{8\varphi_n(r)}{r} + \sqrt{\frac{2V\tau}{nr^{2-\vartheta}}} + \frac{5B\tau}{3nr} \tag{7.17}$$

with probability P^n not less than $1-e^{-\tau}$. Consequently, our next goal is to find functions φ_n satisfying (7.16). To do this, we need the following definitions.

Definition 7.8. *Let $(\Theta, \mathcal{C}, \nu)$ be a probability space and $\varepsilon_i : \Theta \to \{-1, 1\}$, $i = 1, \ldots, n$, be independent random variables with $\nu(\varepsilon_i = 1) = \nu(\varepsilon_i = -1) = 1/2$ for all $i = 1, \ldots, n$. Then $\varepsilon_1, \ldots, \varepsilon_n$ is called a **Rademacher sequence** with respect to ν.*

Statistically speaking, Rademacher sequences model repeated coin flips. Besides this obvious interpretation, they are, however, also an important tool in several branches of pure mathematics.

The following definition uses Rademacher sequences to introduce a new type of expectation of suprema. This new type will then be used to find functions φ_n satisfying (7.16).

Definition 7.9. *Let $\mathcal{H} \subset \mathcal{L}_0(Z)$ be a non-empty set and $\varepsilon_1, \ldots, \varepsilon_n$ be a Rademacher sequence with respect to some distribution ν. Then, for $D := (z_1, \ldots, z_n) \in Z^n$, the n-th* ***empirical Rademacher average of*** *$\mathcal{H}$ is defined by*

$$\mathrm{Rad}_D(\mathcal{H}, n) := \mathbb{E}_\nu \sup_{h \in \mathcal{H}} \left| \frac{1}{n} \sum_{i=1}^{n} \varepsilon_i h(z_i) \right| . \tag{7.18}$$

Note that in the preceding definition we do not need to ensure that the supremum on the right-hand side of (7.18) is measurable since the expectation $\mathbb{E}_\nu$ reduces to a finite sum over 2^n summands.

Some simple structural properties of empirical Rademacher averages can be found in Exercise 7.2, whereas more advanced properties are collected in Section A.8. For now we need the following result, which follows directly from Corollary A.8.2.

Proposition 7.10 (Symmetrization). *Let $\mathcal{H} \subset \mathcal{L}_\infty(Z)$ be a separable Carathéodory set with $\sup_{h \in \mathcal{H}} \|h\|_\infty < \infty$ and P be a distribution on Z. Then for all $n \geq 1$ we have*

$$\mathbb{E}_{D \sim \mathrm{P}^n} \sup_{h \in \mathcal{H}} \left| \mathbb{E}_\mathrm{P} h - \mathbb{E}_\mathrm{D} h \right| \;\leq\; 2\, \mathbb{E}_{D \sim \mathrm{P}^n} \mathrm{Rad}_D(\mathcal{H}, n) . \tag{7.19}$$

Combining the preceding proposition with (7.16), we see that it suffices to bound $\mathbb{E}_{D \sim \mathrm{P}^n} \mathrm{Rad}_D(\mathcal{H}_r, n)$. To achieve this, we will first focus on bounding empirical Rademacher averages. We begin with the following lemma, which can be used to bound Rademacher averages of *finite* function classes.

Lemma 7.11. *Let $(\Theta, \mathcal{C}, \nu)$ be a probability space and $g_1, \ldots, g_m \in \mathcal{L}_0(\Theta)$ be functions for which there exists a constant $K > 0$ satisfying $\mathbb{E}_\nu e^{\lambda g_i} \;\leq\; e^{\lambda^2 K^2}$ for all $\lambda \in \mathbb{R}$ and $i = 1, \ldots, m$. Then we have*

$$\mathbb{E}_\nu \max_{1 \leq i \leq m} |g_i| \;\leq\; 2K \sqrt{\ln(2m)} .$$

Proof. Writing $\max_{1 \leq j \leq m} \pm \lambda g_i := \max\{\lambda g_1, \ldots, \lambda g_m, -\lambda g_1, \ldots, -\lambda g_m\}$ for $\lambda > 0$, we obtain

$$\mathbb{E}_\nu \max_{1 \leq i \leq m} |g_i| \;=\; \lambda^{-1} \mathbb{E}_\nu \max_{1 \leq j \leq m} \pm \lambda g_i \;\leq\; \lambda^{-1} \ln\bigl(\mathbb{E}_\nu e^{\max_{1 \leq j \leq m} \pm \lambda g_i}\bigr)$$

by Jensen's inequality. Moreover, by the monotonicity of the exponential function, we have

$$\mathbb{E}_\nu e^{\max_{1 \leq j \leq m} \pm \lambda g_i} \leq \mathbb{E}_\nu \max_{1 \leq j \leq m} e^{\pm \lambda g_i} \leq \sum_{j=1}^{m} \mathbb{E}_\nu \bigl(e^{\lambda g_i} + e^{-\lambda g_i}\bigr) \leq 2m\, e^{\lambda^2 K^2} .$$

Combining both estimates, we thus find

$$\mathbb{E}_\nu \max_{1 \leq i \leq m} |g_i| \leq \lambda^{-1} \ln\bigl(2m\, e^{\lambda^2 K^2}\bigr) = \lambda^{-1} \ln(2m) + \lambda K^2$$

for all $\lambda > 0$. For $\lambda := K^{-1} \sqrt{\ln(2m)}$, we then obtain the assertion. □

The preceding lemma only provides a bound for suprema over *finitely* many functions. However, if we assume that we have a set of functions $\mathcal{H}$ that can be suitably approximated by a *finite* set of functions, then it seems natural to ask whether we can also bound the supremum over this possibly infinite set $\mathcal{H}$. The following theorem gives a positive answer to this question with the help of entropy numbers.

Theorem 7.12 (Dudley's chaining). *Let (T,d) be a separable metric space, $(\Theta,\mathcal{C},\nu)$ be a probability space, and $(g_t)_{t\in T}\subset\mathcal{L}_0(\Theta)$ be a Carathéodory family. Assume that there exist a $t_0\in T$ and a constant $K>0$ such that $g_{t_0}=0$ and*

$$\mathbb{E}_\nu\, e^{\lambda(g_s-g_t)} \;\le\; e^{\lambda^2K^2\cdot d^2(s,t)}\,, \qquad\qquad s,t\in T,\,\lambda\in\mathbb{R}. \tag{7.20}$$

Then we have

$$\mathbb{E}_\nu \sup_{t\in T}|g_t| \;\le\; 2\sqrt{\ln 4}\,K\biggl(\sum_{i=1}^\infty 2^{i/2}e_{2^i}(T,d)+\sup_{t\in T}d(t,t_0)\biggr).$$

Proof. Without loss of generality, we may assume that all entropy numbers are finite, i.e., $e_n(T,d)<\infty$ for all $n\ge 1$. We fix an $\varepsilon>0$. For $i\ge 1$, we define $s_{2^i}:=(1+\varepsilon)e_{2^i}(T,d)$, and in addition we write $s_1:=\sup_{t\in T}d(t,t_0)$. For $i\ge 1$, we further fix an s_{2^i}-net T_i of T such that $|T_i|\le 2^{2^i-1}$. In addition, we write $T_0:=\{t_0\}$. Then there exist maps $\pi_i:T\to T_i$, $i\ge 0$, such that $d(t,\pi_i(t))\le s_{2^i}$ for all $t\in T$ and $\pi_i(t)=t$ for all $t\in T_i$. Let us now fix a $j\ge 0$. We define maps $\gamma_i:T\to T_i$, $i=0,\ldots,j$, recursively by $\gamma_j:=\pi_j$ and $\gamma_{i-1}:=\pi_{i-1}\circ\gamma_i$, $i=j,\ldots,1$. Note that $T_0=\{t_0\}$ implies $\gamma_0(t)=t_0$ for all $t\in T$. For $k\ge j$ and $t\in T_j$, we hence obtain

$$|g_t|=|g_t-g_{t_0}|=\biggl|\sum_{i=1}^j\bigl(g_{\gamma_i(t)}-g_{\gamma_{i-1}(t)}\bigr)\biggr|\le\sum_{i=1}^k\max_{t\in T_i}\bigl|g_t-g_{\pi_{i-1}(t)}\bigr|\,,$$

which for $S_k:=T_0\cup\cdots\cup T_k$ implies

$$\max_{t\in S_k}|g_t|\le\sum_{i=1}^k\max_{t\in T_i}\bigl|g_t-g_{\pi_{i-1}(t)}\bigr|\,.$$

Moreover, (7.20) yields

$$\mathbb{E}_\nu e^{\lambda(g_t-g_{\pi_{i-1}(t)})}\le e^{\lambda^2K^2s_{2^{i-1}}^2}$$

for all $t\in T_i$, $i\ge 1$, and $\lambda\in\mathbb{R}$. Consequently, Lemma 7.11 implies

$$\mathbb{E}_\nu\max_{t\in S_k}|g_t|\le\sum_{i=1}^k\mathbb{E}_\nu\max_{t\in T_i}\bigl|g_t-g_{\pi_{i-1}(t)}\bigr|\le 2K\sum_{i=1}^k\sqrt{\ln(2\cdot 2^{2^i-1})}\,s_{2^{i-1}}\,.$$

Let us write $S:=\bigcup_{i=0}^\infty T_i$. Then we have $\max_{t\in S_k}|g_t|\nearrow\sup_{t\in S}|g_t|$ for $k\to\infty$, and by Beppo Levi's theorem we hence obtain

$$\mathbb{E}_\nu \sup_{t\in S} |g_t| \;\le\; 2\sqrt{\ln 4}\, K \sum_{i=0}^{\infty} 2^{i/2} s_{2^i}\,.$$

Since S is dense in T, we then obtain the assertion by (7.13), the definition of s_{2^i}, and taking the limit $\varepsilon \to 0$. □

With the help of Dudley's chaining, we can now establish our first bound on empirical Rademacher averages.

Theorem 7.13. *For every non-empty set $\mathcal{H} \subset \mathcal{L}_0(Z)$ and every finite sequence $D := (z_1, \ldots, z_n) \in Z^n$, we have*

$$\mathrm{Rad}_D(\mathcal{H}, n) \le \sqrt{\frac{\ln 16}{n}} \left(\sum_{i=1}^{\infty} 2^{i/2}\, e_{2^i}\big(\mathcal{H} \cup \{0\}, L_2(\mathrm{D})\big) + \sup_{h\in\mathcal{H}} \|h\|_{L_2(\mathrm{D})} \right).$$

Proof. We consider $T := \mathcal{H} \cup \{0\}$ as a subset of $L_2(\mathrm{D})$ and equip it with the metric d defined by $\|\cdot\|_{L_2(\mathrm{D})}$. Since $L_2(\mathrm{D})$ is finite-dimensional, (T, d) is a separable metric space. Let us now fix a Rademacher sequence $\varepsilon_1, \ldots, \varepsilon_n$ with respect to a distribution ν on some Θ. Moreover, for $h \in T$, we define the function $g_h : \Theta \to \mathbb{R}$ by $g_h(\theta) := \frac{1}{n}\sum_{i=1}^n \varepsilon_i(\theta) h(z_i)$, $\theta \in \Theta$. Obviously, this yields $g_0 = 0$ and $g_h \in \mathcal{L}_0(\Theta)$, $h \in T$. Moreover, the convergence with respect to $\|\cdot\|_{L_2(\mathrm{D})}$ implies pointwise convergence on $z_1, \ldots, z_n$, and hence, for each fixed $\theta \in \Theta$, the map $h \mapsto g_h(\theta)$ is continuous with respect to the metric d. In other words, $(g_h)_{h\in T}$ is a separable Carathéodory family. Let us now establish a bound of the form (7.20). To this end, we observe that for $a \in \mathbb{R}$ and $i = 1, \ldots, n$ we have $\mathbb{E}_\nu e^{a\varepsilon_i} = \frac{1}{2}(e^a + e^{-a}) \le e^{a^2/2}$. For $h \in L_2(\mathrm{D})$, the independence of $\varepsilon_1, \ldots, \varepsilon_n$ and $\|h\|^2_{L_2(\mathrm{D})} = \frac{1}{n}\sum_{i=1}^n h^2(z_i)$ thus yields

$$\mathbb{E}_\nu \exp\left(\frac{\lambda}{n}\sum_{i=1}^n \varepsilon_i h(z_i)\right) = \prod_{i=1}^n \mathbb{E}_\nu e^{\frac{\lambda h(z_i)}{n}\varepsilon_i} \le \prod_{i=1}^n e^{\frac{\lambda^2 h^2(z_i)}{2n^2}} = \exp\left(\frac{\lambda^2 \|h\|^2_{L_2(\mathrm{D})}}{2n}\right)$$

for all $\lambda \in \mathbb{R}$. From this we conclude that

$$\mathbb{E}_\nu \exp\big(\lambda(g_{h_1} - g_{h_2})\big) \le \exp\left(\frac{\lambda^2 \|h_1 - h_2\|^2_{L_2(\mathrm{D})}}{2n}\right)$$

for all $\lambda \in \mathbb{R}$ and all $h_1, h_2 \in T$. Consequently, (7.20) is satisfied for $K := (2n)^{-1/2}$, and hence we obtain the assertion by Theorem 7.12. □

Before we return to our main goal, i.e., estimating $\mathbb{E}_{D\sim \mathrm{P}^n} \mathrm{Rad}_D(\mathcal{H}_r, n)$ for $\mathcal{H}_r$ defined by (7.15), we need two simple lemmas. The first one compares the entropy numbers of $\mathcal{H} \cup \{0\}$ with those of $\mathcal{H}$.

Lemma 7.14. *For all $\mathcal{H} \subset \mathcal{L}_0(Z)$, all $D \in Z^n$, and all $i \ge 2$, we have*

$$\begin{aligned} e_1\big(\mathcal{H} \cup \{0\}, L_2(\mathrm{D})\big) &\le \sup_{h\in\mathcal{H}} \|h\|_{L_2(\mathrm{D})}\,, \\ e_i\big(\mathcal{H} \cup \{0\}, L_2(\mathrm{D})\big) &\le e_{i-1}\big(\mathcal{H}, L_2(\mathrm{D})\big)\,. \end{aligned}$$

Proof. The first inequality immediately follows from $0 \in \mathcal{H} \cup \{0\}$. To show the second inequality, we fix an ε-net $T \subset \mathcal{H}$ of $\mathcal{H}$ having cardinality $|T| \le 2^{i-2}$. Then $T \cup \{0\} \subset \mathcal{H} \cup \{0\}$ is an ε-net of $\mathcal{H} \cup \{0\}$ satisfying $|T \cup \{0\}| \le 2^{i-2} + 1$. Using $2^{i-2} + 1 \le 2^{i-1}$, $i \ge 2$, we then find the second assertion. □

The second technical lemma estimates sums of the form $\sum_{i=1}^{\infty} 2^{i/2} s_{2^i}$ for positive sequences (s_i) of known decay.

Lemma 7.15. *Let $(s_i) \subset [0, \infty)$ be a decreasing sequence for which there are $a > 0$ and $p \in (0, 1)$ such that $s_1 \le a\, 2^{-\frac{1}{2p}}$ and $s_i \le a\, i^{-\frac{1}{2p}}$ for all $i \ge 2$. Then we have*

$$\sum_{i=0}^{\infty} 2^{i/2} s_{2^i} \;\le\; \frac{\sqrt{2}\, C_p^p}{(\sqrt{2}-1)(1-p)}\, a^p\, s_1^{1-p}\,, \tag{7.21}$$

where

$$C_p := \frac{\sqrt{2}-1}{\sqrt{2} - 2^{\frac{2p-1}{2p}}} \cdot \frac{1-p}{p}\,. \tag{7.22}$$

Proof. Let us fix a $t \ge 0$ and a natural number $n \ge 1$ such that $n - 1 \le t < n$. Then a simple application of the geometric series yields

$$\sum_{i=0}^{n-1} 2^{i/2} s_{2^i} \le s_1 \sum_{i=0}^{n-1} 2^{i/2} = s_1\, \frac{2^{n/2} - 1}{\sqrt{2} - 1} \le \frac{\sqrt{2}\, s_1}{\sqrt{2} - 1}\, 2^{t/2}\,,$$

and a similar argument shows that

$$\sum_{i=n}^{\infty} 2^{i/2} s_{2^i} \le a \sum_{i=n}^{\infty} 2^{\frac{i}{2} - \frac{i}{2p}} = a\biggl(\sum_{i=0}^{\infty} 2^{\frac{i}{2} - \frac{i}{2p}} - \sum_{i=0}^{n-1} 2^{\frac{i}{2} - \frac{i}{2p}} \biggr) \le \frac{a\, 2^{\frac{t(p-1)}{2p}}}{1 - 2^{\frac{p-1}{2p}}}\,.$$

Combining both estimates, we then obtain that for all $t \ge 0$ we have

$$\sum_{i=0}^{\infty} 2^{i/2} s_{2^i} \le c_1 e^{\alpha_1 t} + c_2 e^{\alpha_2 t}\,, \tag{7.23}$$

where $c_1 := \frac{\sqrt{2}}{\sqrt{2}-1} s_1$, $\alpha_1 := \frac{\ln 2}{2}$, $c_2 := a\bigl(1 - 2^{\frac{p-1}{2p}}\bigr)^{-1}$, and $\alpha_2 := \frac{(p-1)\ln 2}{2p}$. Now it is easy to check that the function $t \mapsto c_1 e^{\alpha_1 t} + c_2 e^{\alpha_2 t}$ is minimized at

$$t^* := \frac{1}{\alpha_2 - \alpha_1} \ln\biggl(-\frac{\alpha_1 c_1}{\alpha_2 c_2} \biggr) = \frac{2p}{\ln 2} \ln\biggl(\frac{a C_p}{s_1} \biggr) \ge \frac{2p}{\ln 2} \ln\bigl(2^{\frac{1}{2p}} C_p \bigr)\,.$$

In order to show $t^* \ge 0$, we write $f(x) := (\sqrt{2} - 1) 2^x x$ and $g(x) = 2^{x/2} - 1$ for $x \ge 0$. Since for $x_p := \frac{1}{p} - 1$ we have

$$2^{\frac{1}{2p}} C_p = (\sqrt{2} - 1) \cdot \frac{2^{(\frac{1}{p} - 1)\frac{1}{2}}}{1 - 2^{-(\frac{1}{p} - 1)\frac{1}{2}}} \cdot \biggl(\frac{1}{p} - 1 \biggr) = \frac{f(x_p)}{g(x_p)}\,,$$

it then suffices to show $f(x) \geq g(x)$ for all $x > 0$. However, the latter follows from $f(0) = g(0) = 0$,

$$f'(x) = (\sqrt{2}-1)2^x(x\ln 2 + 1) \geq 2^x(\sqrt{2}-1) > 2^{x/2}\ln\sqrt{2} = g'(x)\,, \qquad x \geq 0,$$

and the fundamental theorem of calculus. Plugging t^* into (7.23) together with some simple but tedious calculations then yields the assertion. □

With the help of the preceding lemmas, we can now establish an upper bound for expectations of empirical Rademacher averages.

Theorem 7.16. *Let $\mathcal{H} \subset \mathcal{L}_0(Z)$ be a separable Carathéodory set and* P *be a distribution on Z. Suppose that there exist constants $B \geq 0$ and $\sigma \geq 0$ such that $\|h\|_\infty \leq B$ and $\mathbb{E}_{\mathrm{P}} h^2 \leq \sigma^2$ for all $h \in \mathcal{H}$. Furthermore, assume that for a fixed $n \geq 1$ there exist constants $p \in (0,1)$ and $a \geq B$ such that*

$$\mathbb{E}_{D\sim \mathrm{P}^n} e_i(\mathcal{H}, L_2(\mathrm{D})) \;\leq\; a\, i^{-\frac{1}{2p}}\,, \qquad\qquad i \geq 1. \tag{7.24}$$

Then there exist constants $C_1(p) > 0$ and $C_2(p) > 0$ depending only on p such that

$$\mathbb{E}_{D\sim \mathrm{P}^n}\mathrm{Rad}_D(\mathcal{H}, n) \leq \max\Bigl\{C_1(p)\, a^p \sigma^{1-p} n^{-\frac{1}{2}},\, C_2(p)\, a^{\frac{2p}{1+p}} B^{\frac{1-p}{1+p}} n^{-\frac{1}{1+p}}\Bigr\}\,.$$

Proof. For $s_i := \mathbb{E}_{D\sim \mathrm{P}^n} e_i(\mathcal{H}\cup\{0\}, L_2(\mathrm{D}))$, $i \geq 2$, and $\tilde{a} := 2^{\frac{1}{2p}} a$, Lemma 7.14 yields

$$s_i \leq \mathbb{E}_{D\sim \mathrm{P}^n} e_{i-1}(\mathcal{H}, L_2(\mathrm{D})) \leq a(i-1)^{-\frac{1}{2p}} \leq \tilde{a} i^{-\frac{1}{2p}}\,. \tag{7.25}$$

Furthermore, for

$$\delta_D := \sup_{h\in\mathcal{H}} \|h\|_{L_2(\mathrm{D})}\,, \qquad\qquad D \in Z^n\,,$$

and $s_1 := \mathbb{E}_{D\sim \mathrm{P}^n}\delta_D$, we have $s_1 \leq B \leq a = \tilde{a}2^{-\frac{1}{2p}}$. Moreover, the monotonicity of the entropy numbers together with Lemma 7.14 shows that $s_2 \leq \mathbb{E}_{D\sim \mathrm{P}^n} e_1(\mathcal{H}\cup\{0\}, L_2(\mathrm{D})) \leq s_1$, and hence it is easy to conclude that (s_i) is decreasing. By Theorem 7.13 and Lemma 7.15, we hence find

$$\begin{aligned}\mathbb{E}_{D\sim \mathrm{P}^n}\mathrm{Rad}_D(\mathcal{H}, n) &\leq \sqrt{\frac{\ln 16}{n}} \sum_{i=0}^{\infty} 2^{i/2}\, s_{2^i} \\ &\leq \frac{1}{\sqrt{n}} \cdot \frac{2\sqrt{\ln 16}}{(\sqrt{2}-1)(1-p)} C_p^p \left(\mathbb{E}_{D\sim \mathrm{P}^n}\delta_D\right)^{1-p} a^p\,.\end{aligned}$$

Moreover, Corollary A.8.5 yields

$$\mathbb{E}_{D\sim \mathrm{P}^n}\delta_D \leq (\mathbb{E}_{D\sim \mathrm{P}^n}\delta_D^2)^{1/2} \leq \left(\sigma^2 + 8B\,\mathbb{E}_{D\sim \mathrm{P}^n}\mathrm{Rad}_D(\mathcal{H}, n)\right)^{1/2}\,,$$

and hence we obtain

$$\mathbb{E}_{D\sim \mathrm{P}^n}\mathrm{Rad}_D(\mathcal{H},n) \leq \frac{2\sqrt{\ln 16}\,C_p^p}{(\sqrt{2}-1)(1-p)}\cdot\frac{a^p}{\sqrt{n}}\,\big(\sigma^2+8B\,\mathbb{E}_{D\sim \mathrm{P}^n}\mathrm{Rad}_D(\mathcal{H},n)\big)^{\frac{1-p}{2}}\,.$$

In the case $\sigma^2 \geq 8B\,\mathbb{E}_{D\sim \mathrm{P}^n}\mathrm{Rad}_D(\mathcal{H},n)$, we conclude that

$$\mathbb{E}_{D\sim \mathrm{P}^n}\mathrm{Rad}_D(\mathcal{H},n) \leq \frac{2\sqrt{\ln 256}\,C_p^p}{(\sqrt{2}-1)(1-p)2^{p/2}}\cdot\frac{a^p}{\sqrt{n}}\cdot\sigma^{1-p}\,,$$

and in the case $\sigma^2 < 8B\,\mathbb{E}_{D\sim \mathrm{P}^n}\mathrm{Rad}_D(\mathcal{H},n)$ we have

$$\mathbb{E}_{D\sim \mathrm{P}^n}\mathrm{Rad}_D(\mathcal{H},n) \leq \frac{8\sqrt{\ln 16}\,C_p^p}{(\sqrt{2}-1)(1-p)4^p}\cdot\frac{a^p}{\sqrt{n}}\cdot\big(B\,\mathbb{E}_{D\sim \mathrm{P}^n}\mathrm{Rad}_D(\mathcal{H},n)\big)^{\frac{1-p}{2}}\,.$$

Simple algebraic transformations then yield the assertion. □

Numerical calculations show that the constants obtained in the proof of Theorem 7.16 satisfy $(1-p)C_1(p) \in [6.8, 12.25]$ and $(1-p)C_2(p) \in [5.68, 6.8]$. Further calculations for $C_1(p)$ yield $\lim_{p\to 0} C_1(p) \leq 11.38$ and $\lim_{p\to 1}(1-p)$ $C_1(p) \approx 6.8$. Finally, similar calculations for $C_2(p)$ give $\lim_{p\to 0} C_2(p) \approx 5.68$ and $\lim_{p\to 1}(1-p)C_2(p) \leq 6.8$. However, it is obvious from the proof above that we can, e.g., obtain smaller values for $C_2(p)$ for the price of larger values for $C_1(p)$, and hence the calculations above only illustrate how $C_1(p)$ and $C_2(p)$ can be estimated.

Before we use these estimates on the Rademacher averages to bound the term

$$\mathbb{E}_{D\sim \mathrm{P}^n}\sup_{h\in\mathcal{H}_r}\big|\mathbb{E}_{\mathrm{P}}h - \mathbb{E}_{\mathrm{D}}h\big|$$

on the left-hand side of (7.16), we present a simple lemma that estimates the entropy numbers of a set $\mathcal{H} := \{L\circ f - L\circ f^*_{L,\mathrm{P}} : f \in \mathcal{F}\}$ by the entropy numbers of the "base" set $\mathcal{F}$.

Lemma 7.17. *Let $L : X \times Y \times \mathbb{R} \to [0,\infty)$ be a locally Lipschitz continuous loss, P be a distribution on $X \times Y$, and $q \in [1,\infty]$ and $M > 0$ be constants. Assume that we have $\mathbb{E}_{(x,y)\sim \mathrm{P}}L^q(x,y,0) < \infty$. Furthermore, let $\mathcal{F} \subset \mathcal{L}_q(\mathrm{P}_X)$ be a non-empty set and $f_0 \in \mathcal{L}_0(\mathrm{P}_X)$ be an arbitrary function. We write*

$$\mathcal{H} := \big\{L\circ \widehat{f} - L\circ \widehat{f_0} : f \in \mathcal{F}\big\}\,,$$

where $\widehat{\cdot}$ is the clipping operation at M defined in (2.14). Then we have

$$e_n\big(\mathcal{H}, \|\cdot\|_{L_q(\mathrm{P})}\big) \;\leq\; |L|_{M,1}\cdot e_n\big(\mathcal{F}, \|\cdot\|_{L_q(\mathrm{P}_X)}\big)\,, \qquad n \geq 1.$$

Proof. For $L\circ\widehat{\mathcal{F}} := \{L\circ\widehat{f} : f \in \mathcal{F}\}$, we have $L\circ\widehat{\mathcal{F}} \subset \mathcal{L}_q(\mathrm{P})$ by (2.11) and $\mathbb{E}_{(x,y)\sim \mathrm{P}}L^q(x,y,0) < \infty$. In addition, we obviously have

$$e_n\big(\mathcal{H}, \|\cdot\|_{L_q(\mathrm{P})}\big) \;\leq\; e_n\big(L\circ\widehat{\mathcal{F}}, \|\cdot\|_{L_q(\mathrm{P})}\big)\,, \qquad n \geq 1.$$

Now let $f_1, \ldots, f_{2^{n-1}}$ be an ε-net of $\mathcal{F}$ with respect to $\|\cdot\|_{L_q(\mathrm{P}_X)}$. For $f \in \mathcal{F}$, there then exists an $i \in \{1, \ldots, 2^{n-1}\}$ such that $\|f - f_i\|_{L_q(\mathrm{P}_X)} \leq \varepsilon$. Moreover, the clipping operation obviously satisfies $|\widehat{s} - \widehat{t}| \leq |s - t|$ for all $s, t \in \mathbb{R}$. For $q < \infty$, we thus obtain

$$\begin{aligned}
\|L \circ \widehat{f} - L \circ \widehat{f}_i\|^q_{L_q(\mathrm{P})} &= \int_{X \times Y} \bigl|L(x, y, \widehat{f}(x)) - L(x, y, \widehat{f}_i(x))\bigr|^q \, d\mathrm{P}(x, y) \\
&\leq |L|^q_{M,1} \int_X \bigl|\widehat{f}(x) - \widehat{f}_i(x)\bigr|^q \, d\mathrm{P}_X(x) \\
&\leq |L|^q_{M,1} \cdot \varepsilon^q \,,
\end{aligned}$$

and consequently, $L \circ \widehat{f}_1, \ldots, L \circ \widehat{f}_{2^{n-1}}$ is an $|L|_{M,1} \cdot \varepsilon$-net of $L \circ \widehat{\mathcal{F}}$ with respect to $\|\cdot\|_{L_q(\mathrm{P})}$. From this we easily find the assertion. The case $q = \infty$ can be shown analogously. □

Let us now return to our example at the beginning of this section, where we applied Talagrand's inequality to the set $\mathcal{G}_r := \{g_{f,r} : f \in \mathcal{F}\}$ defined by (7.12). To this end, let us assume that L is a locally Lipschitz continuous loss function[3] and $\mathcal{F} \subset \mathcal{L}_0(X)$ is a non-empty subset. Assume that $\mathcal{F}$ is equipped with a complete, separable metric dominating the pointwise convergence and that all $f \in \mathcal{F}$ satisfy $\|f\|_\infty \leq M$ for a suitable constant $M > 0$. Moreover, we assume that there exists a $B > 0$ such that

$$L(x, y, t) \leq B \,, \qquad (x, y) \in X \times Y, \, t \in [-M, M].$$

In addition, let P be a distribution on $X \times Y$ and $f^*_{L,\mathrm{P}} : X \to [-M, M]$ be a Bayes decision function. Note that combining these assumptions, we see that the supremum bound (7.5) of Theorem 7.2 holds for all $f \in \mathcal{F}$. Furthermore, these assumption match those of Proposition 6.22, which established our first oracle inequality for ERM over infinite function classes. Assume further that there exist constants $\vartheta \in [0, 1]$ and $V \geq B^{2-\vartheta}$ such that for all $f \in \mathcal{F}$ we have

$$\mathbb{E}_\mathrm{P}\bigl(L \circ f - L \circ f^*_{L,\mathrm{P}}\bigr)^2 \leq V \cdot \bigl(\mathbb{E}_\mathrm{P}(L \circ f - L \circ f^*_{L,\mathrm{P}})\bigr)^\vartheta \,. \tag{7.26}$$

Let us now define $\mathcal{H}_r$ by (7.15), i.e., $\mathcal{H}_r := \{h_f : f \in \mathcal{F} \text{ and } \mathbb{E}_\mathrm{P} h_f \leq r\}$, where $h_f := L \circ f - L \circ f^*_{L,\mathrm{P}}$. Finally, let us assume that there exist constants $a > 0$ and $p \in (0, 1)$ such that

$$\mathbb{E}_{D \sim \mathrm{P}^n} e_i(\mathcal{F}, L_2(\mathrm{D})) \;\leq\; a \, i^{-\frac{1}{2p}} \,, \qquad i, n \geq 1. \tag{7.27}$$

Now note that $\|f\|_\infty \leq M$ implies $\widehat{f} = f$ for all $f \in \mathcal{F}$, where $\widehat{\cdot}$ denotes the clipping operation at M. Applying Lemma 7.17 together with (A.36) therefore yields $\mathbb{E}_{D \sim \mathrm{P}^n} e_i(\mathcal{H}_r, L_2(\mathrm{D})) \leq 2a|L|_{M,1} i^{-\frac{1}{2p}}$ for all $i, n \geq 1$ and all

[3] Loss functions that can be clipped will be considered in the following section. Here we assume for the sake of simplicity that clipping is superfluous.

$r > r^* := \inf\{\mathbb{E}_P h_f : f \in \mathcal{F}\}$. In addition, (7.26) implies $\mathbb{E}_P h^2 \le V r^\vartheta$ for all $h \in \mathcal{H}_r$, and thus we obtain

$$\mathbb{E}_{D\sim P^n} \mathrm{Rad}_D(\mathcal{H}_r, n) \le C \max\left\{ r^{\frac{\vartheta(1-p)}{2}} n^{-\frac{1}{2}}, n^{-\frac{1}{1+p}} \right\}$$

by Theorem 7.16, where the constant C depends on a, p, M, B, and V. By (7.17) and Proposition 7.10, we thus find that

$$\sup_{f\in\mathcal{F}} \frac{\mathbb{E}_P h_f - \mathbb{E}_D h_f}{\mathbb{E}_P h_f + r} < \frac{16C}{\sqrt{n r^{2-\vartheta(1-p)}}} + \frac{16C}{n^{1/(1+p)} r} + \sqrt{\frac{2V\tau}{n r^{2-\vartheta}}} + \frac{5B\tau}{3nr}$$

holds with probability P^n not less than $1 - e^{-\tau}$. Let us now assume that we consider a measurable ERM over the set $\mathcal{F}$. By replacing (7.9) with the inequality above, the proof of Theorem 7.2 then yields after some basic yet exhausting calculations[4] that for fixed $\tau \ge 1$ and $n \ge 1$ we have

$$\mathcal{R}_{L,P}(f_D) - \mathcal{R}^*_{L,P} \;\le\; 6\big(\mathcal{R}^*_{L,P,\mathcal{F}} - \mathcal{R}^*_{L,P}\big) + c\sqrt{\tau} n^{-\frac{1}{2-\vartheta+\vartheta p}} \tag{7.28}$$

with probability P^n not less than $1 - e^{-\tau}$, where c is a constant independent of τ and n. To appreciate this oracle inequality, we briefly compare it with our previous results for ERM. Our first approach using covering numbers led to Proposition 6.22, which for classes $\mathcal{F} \subset \mathcal{L}_\infty(X)$ satisfying the assumption

$$e_i(\mathcal{F}, \|\cdot\|_\infty) \;\le\; a\, i^{-\frac{1}{2p}}, \qquad\qquad i \ge 1, \tag{7.29}$$

showed that for fixed $\tau \ge 1$ and $n \ge 1$ we have

$$\mathcal{R}_{L,P}(f_D) - \mathcal{R}^*_{L,P} \;\le\; (\mathcal{R}^*_{L,P,\mathcal{F}} - \mathcal{R}^*_{L,P}) + \tilde{c}\sqrt{\tau} n^{-\frac{1}{2+2p}} \tag{7.30}$$

with probability P^n not less than $1 - e^{-\tau}$, where $\tilde{c}$ is a constant independent of τ and n. Now note that, for $p \in (0,1)$, condition (7.29) implies (7.27) and hence the oracle inequality (7.28) derived by the more involved techniques also holds. Moreover, we have $2 + 2p > 2 - \vartheta + \vartheta p$ by at least $2p$, and consequently (7.30) has worse behavior in the sample size n than (7.28). In this regard, it is also interesting to note that a brute-force approach (see Exercise 7.3) for generalizing Theorem 7.2 with the help of covering numbers does provide an improvement compared with (7.30). However, this improved oracle inequality is still not as sharp as (7.28).

Another difference between the oracle inequalities above is that (7.30) estimates against $\mathcal{R}^*_{L,P,\mathcal{F}} - \mathcal{R}^*_{L,P}$, which at first glance seems to be more interesting than $6(\mathcal{R}^*_{L,P,\mathcal{F}} - \mathcal{R}^*_{L,P})$ considered in (7.28). However, in the following, we are mainly interested in situations where the sets $\mathcal{F}$ become larger with the sample size, and hence this effect will be negligible for our purposes.

Finally, to compare (7.28) with the oracle inequality of Theorem 7.2, we assume that $\mathcal{F}$ is a *finite* set. Then it is easy to see that the entropy assumption (7.27) is satisfied for all $p \in (0,1)$, and hence (7.28) essentially recovers, i.e., modulo an arbitrarily small change in the exponent of n, Theorem 7.2.

[4] Since we will establish a more general inequality in the next section we omit the details of the estimation as well as an explicit upper bound on the constant c.

7.4 Refined Oracle Inequalities for SVMs

The goal of this section is to establish improved oracle inequalities for SVMs using the ideas of the previous section.

Let us begin by recalling Section 6.5, where we saw that oracle inequalities for SVMs can be used to investigate data-dependent parameter selection strategies if the loss can be *clipped*. In the following, we will therefore focus solely on such loss functions. Moreover, we will first consider the following modification of regularized empirical risk minimization.

Definition 7.18. *Let $L : X \times Y \times \mathbb{R} \to [0,\infty)$ be a loss that can be clipped, $\mathcal{F} \subset \mathcal{L}_0(X)$ be a subset, $\Upsilon : \mathcal{F} \to [0,\infty)$ be a function, and $\epsilon \geq 0$. A learning method whose decision functions f_D satisfy*

$$\Upsilon(f_D) + \mathcal{R}_{L,\mathrm{D}}(\widehat{f}_D) \;\leq\; \inf_{f\in\mathcal{F}} \Upsilon(f) + \mathcal{R}_{L,\mathrm{D}}(f) + \epsilon \tag{7.31}$$

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

Moreover, in the case $\epsilon = 0$, we simply speak of ***clipped regularized empirical risk minimization (CR-ERM)****.*

Note that on the right-hand side of (7.31) the unclipped loss is considered, and hence CR-ERM does *not* necessarily minimize the regularized clipped empirical risk $\Upsilon(\,\cdot\,) + \mathcal{R}_{L,\mathrm{D}}(\widehat{\,\cdot\,})$. Moreover, in general CR-ERMs do *not* minimize the regularized risk $\Upsilon(\,\cdot\,) + \mathcal{R}_{L,\mathrm{D}}(\,\cdot\,)$ either, because on the left-hand side of (7.31) the clipped function is considered. However, if we have a minimizer of the *un*clipped regularized risk, then it automatically satisfies (7.31). In particular, SVM decision functions satisfy (7.31) for the regularizer $\Upsilon := \lambda\|\cdot\|_H^2$ and $\epsilon := 0$. In other words, SVMs are CR-ERMs.

Before we establish an oracle inequality for CR-ERMs, let us first ensure that there exist measurable versions. This is done in the following lemma.

Lemma 7.19 (Measurability of CR-ERMs). *Let $L : X \times Y \times \mathbb{R} \to [0,\infty)$ be a loss that can be clipped and $\mathcal{F} \subset \mathcal{L}_0(X)$ be a subset that is equipped with a complete, separable metric dominating the pointwise convergence. Moreover, let $\Upsilon : \mathcal{F} \to [0,\infty)$ be a function that is measurable with respect to the corresponding Borel σ-algebra on $\mathcal{F}$. Then, for all $\epsilon > 0$, there exists a measurable ϵ-CR-ERM with respect to L, $\mathcal{F}$, and Υ. Moreover, if there exists a CR-ERM, then there also exists a measurable CR-ERM. Finally, in both cases, the map $D \mapsto \Upsilon(f_D)$ mapping $(X \times Y)^n$ to $[0,\infty)$ is measurable for all $n \geq 1$.*

Proof. The maps $(D,f) \mapsto \Upsilon(f) + \mathcal{R}_{L,\mathrm{D}}(\widehat{f})$ and $(D,f) \mapsto \Upsilon(f) + \mathcal{R}_{L,\mathrm{D}}(f)$ are measurable by Lemma 2.11, where for the measurability of the first map we consider the clipped version of L, i.e.,

$$\widehat{L}(x,y,t) := L(x,y,\widehat{t}\,), \qquad (x,y) \in X \times Y,\, t \in \mathbb{R}.$$

Consequently, the maps above are also measurable with respect to the universal completion of the σ-algebra on $(X \times Y)^n$. By part *iii)* of Lemma A.3.18, the map $h : (X \times Y)^n \times \mathcal{F} \to \mathbb{R}$ defined by

$$h(D, f) := \Upsilon(f) + \mathcal{R}_{L,\mathrm{D}}(\widehat{f}\,) - \inf_{f' \in \mathcal{F}} \bigl(\Upsilon(f') + \mathcal{R}_{L,\mathrm{D}}(f')\bigr)\,,$$

$D \in (X \times Y)^n$, $f \in \mathcal{F}$, is therefore measurable with respect to the universal completion of the σ-algebra on $(X \times Y)^n$. For $A := (-\infty, \epsilon]$, we now consider the map $F : (X \times Y)^n \to 2^{\mathcal{F}}$ defined by

$$F(D) := \bigl\{f \in \mathcal{F} : h(D, f) \in A\bigr\}\,, \qquad\qquad D \in (X \times Y)^n.$$

Obviously, $F(D)$ contains exactly the functions f satisfying (7.31). Moreover, in the case $\epsilon > 0$, we obviously have $F(D) \neq \emptyset$ for all $D \in (X \times Y)^n$, while in the case $\epsilon = 0$ this follows from the existence of a CR-ERM. Therefore, part *ii)* of Lemma A.3.18 shows that there exists a measurable map $D \mapsto f_D$ such that $f_D \in F(D)$ for all $D \in (X \times Y)^n$. The first two assertions can then be shown by a literal repetition of the proof of Lemma 6.17. The last assertion follows from the measurability of Υ. □

Before we present the first main result of this section, which establishes an oracle inequality for general ϵ-CR-ERMs, we first need to introduce a few more notations. To this end, we assume that L, $\mathcal{F}$, and Υ satisfy the assumptions of Lemma 7.19. Moreover, let P be a distribution on $X \times Y$ such that there exists a Bayes decision function $f^*_{L,\mathrm{P}} : X \to [-M, M]$, where $M > 0$ is the constant at which L can be clipped. For

$$r^* := \inf_{f \in \mathcal{F}} \Upsilon(f) + \mathcal{R}_{L,\mathrm{P}}(\widehat{f}\,) - \mathcal{R}^*_{L,\mathrm{P}} \tag{7.32}$$

and $r > r^*$, we write

$$\mathcal{F}_r := \{f \in \mathcal{F} \,:\, \Upsilon(f) + \mathcal{R}_{L,\mathrm{P}}(\widehat{f}\,) - \mathcal{R}^*_{L,\mathrm{P}} \leq r\}\,, \tag{7.33}$$

$$\mathcal{H}_r := \{L \circ \widehat{f} \, - L \circ f^*_{L,\mathrm{P}} \,:\, f \in \mathcal{F}_r\}\,, \tag{7.34}$$

where, as usual, $L \circ g$ denotes the function $(x, y) \mapsto L(x, y, g(x))$. Furthermore, assume that there exist constants $B > 0$, $\vartheta \in [0, 1]$, and $V \geq B^{2-\vartheta}$ such that

$$L(x, y, t) \leq B\,, \tag{7.35}$$

$$\mathbb{E}_{\mathrm{P}}\bigl(L \circ \widehat{f} \, - L \circ f^*_{L,\mathrm{P}}\bigr)^2 \leq V \cdot \bigl(\mathbb{E}_{\mathrm{P}}(L \circ \widehat{f} \, - L \circ f^*_{L,\mathrm{P}})\bigr)^{\vartheta}\,, \tag{7.36}$$

for all $(x, y) \in X \times Y$, $t \in [-M, M]$, and $f \in \mathcal{F}$. In the following, (7.35) is called a **supremum bound** and (7.36) is called a **variance bound**.

With the help of these notions, we can now formulate the first main result of this section.

Theorem 7.20 (Oracle inequality for CR-ERMs). *Let $L : X \times Y \times \mathbb{R} \to [0,\infty)$ be a continuous loss that can be clipped at $M > 0$ and that satisfies (7.35) for a constant $B > 0$. Moreover, let $\mathcal{F} \subset \mathcal{L}_0(X)$ be a subset that is equipped with a complete, separable metric dominating the pointwise convergence, and let $\Upsilon : \mathcal{F} \to [0,\infty)$ be a continuous function. Given a distribution P on $X \times Y$ that satisfies (7.36), we define r^* and $\mathcal{H}_r$ by (7.32) and (7.34), respectively. Assume that for fixed $n \geq 1$ there exists a $\varphi_n : [0,\infty) \to [0,\infty)$ such that $\varphi_n(4r) \leq 2\varphi_n(r)$ and*

$$\mathbb{E}_{D\sim \mathrm{P}^n} \mathrm{Rad}_D(\mathcal{H}_r, n) \;\leq\; \varphi_n(r) \tag{7.37}$$

for all $r > r^*$. Finally, fix an $f_0 \in \mathcal{F}$ and a $B_0 \geq B$ such that $\|L \circ f_0\|_\infty \leq B_0$. Then, for all fixed $\epsilon \geq 0$, $\tau > 0$, and $r > 0$ satisfying

$$r > \max\left\{30\varphi_n(r), \left(\frac{72V\tau}{n}\right)^{\frac{1}{2-\vartheta}}, \frac{5B_0\tau}{n}, r^*\right\}, \tag{7.38}$$

every measurable ϵ-CR-ERM satisfies

$$\Upsilon(f_D) + \mathcal{R}_{L,\mathrm{P}}(\widehat{f}_D) - \mathcal{R}^*_{L,\mathrm{P}} \;\leq\; 6\big(\Upsilon(f_0) + \mathcal{R}_{L,\mathrm{P}}(f_0) - \mathcal{R}^*_{L,\mathrm{P}}\big) + 3r + 3\epsilon$$

with probability P^n not less than $1 - 3e^{-\tau}$.

Note that the value of r in Theorem 7.20 is only *implicitly* determined by the relation $r > 30\varphi_n(r)$ appearing in (7.38). Of course, for particular types of functions φ_n, this implicit definition can be made explicit, and we will see in the following that for SVMs this is even a relatively easy task.

Proof. As usual, we define $h_f := L \circ f - L \circ f^*_{L,\mathrm{P}}$ for all $f \in \mathcal{L}_0(X)$. By the definition of f_D, we then have

$$\Upsilon(f_D) + \mathbb{E}_{\mathrm{D}} h_{\widehat{f}_D} \leq \Upsilon(f_0) + \mathbb{E}_{\mathrm{D}} h_{f_0} + \epsilon\,,$$

and consequently we obtain

$$\begin{aligned}
&\Upsilon(f_D) + \mathcal{R}_{L,\mathrm{P}}(\widehat{f}_D) - \mathcal{R}^*_{L,\mathrm{P}} \\
&= \Upsilon(f_D) + \mathbb{E}_{\mathrm{P}} h_{\widehat{f}_D} \\
&\leq \Upsilon(f_0) + \mathbb{E}_{\mathrm{D}} h_{f_0} - \mathbb{E}_{\mathrm{D}} h_{\widehat{f}_D} + \mathbb{E}_{\mathrm{P}} h_{\widehat{f}_D} + \epsilon \\
&= (\Upsilon(f_0) + \mathbb{E}_{\mathrm{P}} h_{f_0}) + (\mathbb{E}_{\mathrm{D}} h_{f_0} - \mathbb{E}_{\mathrm{P}} h_{f_0}) + (\mathbb{E}_{\mathrm{P}} h_{\widehat{f}_D} - \mathbb{E}_{\mathrm{D}} h_{\widehat{f}_D}) + \epsilon
\end{aligned} \tag{7.39}$$

for all $D \in (X \times Y)^n$. Let us first bound the term $\mathbb{E}_{\mathrm{D}} h_{f_0} - \mathbb{E}_{\mathrm{P}} h_{f_0}$. To this end, we further split this difference into

$$\mathbb{E}_{\mathrm{D}} h_{f_0} - \mathbb{E}_{\mathrm{P}} h_{f_0} = \big(\mathbb{E}_{\mathrm{D}}(h_{f_0} - h_{\widehat{f}_0}) - \mathbb{E}_{\mathrm{P}}(h_{f_0} - h_{\widehat{f}_0})\big) + \big(\mathbb{E}_{\mathrm{D}} h_{\widehat{f}_0} - \mathbb{E}_{\mathrm{P}} h_{\widehat{f}_0}\big)\,. \tag{7.40}$$

Now observe that $L \circ f_0 - L \circ \widehat{f}_0 \geq 0$ implies $h_{f_0} - h_{\widehat{f}_0} = L \circ f_0 - L \circ \widehat{f}_0 \in [0, B_0]$, and hence we obtain

$$\mathbb{E}_{\mathrm{P}}\big((h_{f_0} - h_{\widehat{f}_0}) - \mathbb{E}_{\mathrm{P}}(h_{f_0} - h_{\widehat{f}_0})\big)^2 \leq \mathbb{E}_{\mathrm{P}}(h_{f_0} - h_{\widehat{f}_0})^2 \leq B_0\, \mathbb{E}_{\mathrm{P}}(h_{f_0} - h_{\widehat{f}_0})\,.$$

Consequently, Bernstein's inequality stated in Theorem 6.12 and applied to the function $h := (h_{f_0} - h_{\widehat{f}_0}) - \mathbb{E}_{\mathrm{P}}(h_{f_0} - h_{\widehat{f}_0})$ shows that

$$\mathbb{E}_{\mathrm{D}}(h_{f_0} - h_{\widehat{f}_0}) - \mathbb{E}_{\mathrm{P}}(h_{f_0} - h_{\widehat{f}_0}) < \sqrt{\frac{2\tau B_0\, \mathbb{E}_{\mathrm{P}}(h_{f_0} - h_{\widehat{f}_0})}{n}} + \frac{2B_0\tau}{3n}$$

holds with probability P^n not less than $1 - e^{-\tau}$. Moreover, using $\sqrt{ab} \leq \frac{a}{2} + \frac{b}{2}$, we find

$$\sqrt{\frac{2\tau B_0\, \mathbb{E}_{\mathrm{P}}(h_{f_0} - h_{\widehat{f}_0})}{n}} \leq \mathbb{E}_{\mathrm{P}}(h_{f_0} - h_{\widehat{f}_0}) + \frac{B_0\tau}{2n}\,,$$

and consequently we have

$$\mathbb{E}_{\mathrm{D}}(h_{f_0} - h_{\widehat{f}_0}) - \mathbb{E}_{\mathrm{P}}(h_{f_0} - h_{\widehat{f}_0}) < \mathbb{E}_{\mathrm{P}}(h_{f_0} - h_{\widehat{f}_0}) + \frac{7B_0\tau}{6n} \tag{7.41}$$

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

In order to bound the remaining term in (7.40), we now recall the proof of Theorem 7.2, where we note that (7.35) implies (7.5). Consequently, (7.8) shows that with probability P^n not less than $1 - e^{-\tau}$ we have

$$\mathbb{E}_{\mathrm{D}} h_{\widehat{f}_0} - \mathbb{E}_{\mathrm{P}} h_{\widehat{f}_0} < \mathbb{E}_{\mathrm{P}} h_{\widehat{f}_0} + \Big(\frac{2V\tau}{n}\Big)^{\frac{1}{2-\vartheta}} + \frac{4B\tau}{3n}\,.$$

By combining this estimate with (7.41) and (7.40), we now obtain that with probability P^n not less than $1 - 2e^{-\tau}$ we have

$$\mathbb{E}_{\mathrm{D}} h_{f_0} - \mathbb{E}_{\mathrm{P}} h_{f_0} < \mathbb{E}_{\mathrm{P}} h_{f_0} + \Big(\frac{2V\tau}{n}\Big)^{\frac{1}{2-\vartheta}} + \frac{4B\tau}{3n} + \frac{7B_0\tau}{6n}\,, \tag{7.42}$$

i.e., we have established a bound on the second term in (7.39).

Let us now consider the case $n < 72\tau$. Then the assumption $B^{2-\vartheta} \leq V$ implies $B < r$. Combining (7.42) with (7.39) and using both $B \leq B_0$ and $\mathbb{E}_{\mathrm{P}} h_{\widehat{f}_D} - \mathbb{E}_{\mathrm{D}} h_{\widehat{f}_D} \leq 2B$ we hence find

$$\begin{aligned}
&\Upsilon(f_D) + \mathcal{R}_{L,\mathrm{P}}(\widehat{f}_D) - \mathcal{R}^*_{L,\mathrm{P}} \\
&\leq \Upsilon(f_0) + 2\mathbb{E}_{\mathrm{P}} h_{f_0} + \Big(\frac{2V\tau}{n}\Big)^{\frac{1}{2-\vartheta}} + \frac{5B_0\tau}{2n} + (\mathbb{E}_{\mathrm{P}} h_{\widehat{f}_D} - \mathbb{E}_{\mathrm{D}} h_{\widehat{f}_D}) + \epsilon \\
&\leq 6\big(\Upsilon(f_0) + \mathcal{R}_{L,\mathrm{P}}(f_0) - \mathcal{R}^*_{L,\mathrm{P}}\big) + 3r + \varepsilon
\end{aligned}$$

with probability P^n not less than $1 - 2e^{-\tau}$.

Consequently, it remains to consider the case $n \geq 72\tau$. In order to establish a non-trivial bound on the term $\mathbb{E}_{\mathrm{P}} h_{\widehat{f}_D} - \mathbb{E}_{\mathrm{D}} h_{\widehat{f}_D}$ in (7.39), we define functions

$$g_{f,r} := \frac{\mathbb{E}_{\mathrm{P}} h_{\widehat{f}} - h_{\widehat{f}}}{\Upsilon(f) + \mathbb{E}_{\mathrm{P}} h_{\widehat{f}} + r}, \qquad f \in \mathcal{F},\ r > r^*.$$

Obviously, for $f \in \mathcal{F}$, we then have $\|g_{f,r}\|_\infty \le 2Br^{-1}$, and for $\vartheta > 0$, $q := \frac{2}{2-\vartheta}$, $q' := \frac{2}{\vartheta}$, $a := r$, and $b := \mathbb{E}_{\mathrm{P}} h_{\widehat{f}} \neq 0$, the second inequality of Lemma 7.1 yields

$$\mathbb{E}_{\mathrm{P}} g_{f,r}^2 \le \frac{\mathbb{E}_{\mathrm{P}} h_{\widehat{f}}^2}{(\mathbb{E}_{\mathrm{P}} h_{\widehat{f}} + r)^2} \le \frac{(2-\vartheta)^{2-\vartheta}\vartheta^\vartheta\, \mathbb{E}_{\mathrm{P}} h_{\widehat{f}}^2}{4r^{2-\vartheta}(\mathbb{E}_{\mathrm{P}} h_{\widehat{f}})^\vartheta} \le V r^{\vartheta-2}. \tag{7.43}$$

Moreover, for $\vartheta > 0$ and $\mathbb{E}_{\mathrm{P}} h_{\widehat{f}} = 0$, we have $\mathbb{E}_{\mathrm{P}} h_{\widehat{f}}^2 = 0$ by the variance bound (7.36), which in turn implies $\mathbb{E}_{\mathrm{P}} g_{f,r}^2 \le V r^{\vartheta-2}$. Finally, it is not hard to see that $\mathbb{E}_{\mathrm{P}} g_{f,r}^2 \le V r^{\vartheta-2}$ also holds for $\vartheta = 0$. By simple modifications of Lemma 7.6 and its proof, we further see that all families of maps considered below are Carathéodory families. Symmetrization by Proposition 7.10 thus yields

$$\mathbb{E}_{D\sim \mathrm{P}^n} \sup_{f\in\mathcal{F}_r} \bigl|\mathbb{E}_{\mathrm{D}}(\mathbb{E}_{\mathrm{P}} h_{\widehat{f}} - h_{\widehat{f}})\bigr| \le 2\mathbb{E}_{D\sim \mathrm{P}^n} \mathrm{Rad}_D(\mathcal{H}_r, n) \le 2\varphi_n(r).$$

Peeling by Theorem 7.7 together with $\mathcal{F}_r = \{f \in \mathcal{F} : \Upsilon(f) + \mathbb{E}_{\mathrm{P}} h_{\widehat{f}} \le r\}$ hence gives

$$\mathbb{E}_{D\sim \mathrm{P}^n} \sup_{f\in\mathcal{F}} \bigl|\mathbb{E}_{\mathrm{D}} g_{f,r}\bigr| \ \le\ \frac{8\varphi_n(r)}{r}.$$

By Talagrand's inequality in the form of Theorem 7.5 applied to $\gamma := 1/4$, we therefore obtain

$$\mathrm{P}^n\left(D \in (X\times Y)^n : \sup_{f\in\mathcal{F}} \mathbb{E}_{\mathrm{D}} g_{f,r} < \frac{10\varphi_n(r)}{r} + \sqrt{\frac{2V\tau}{nr^{2-\vartheta}}} + \frac{28B\tau}{3nr}\right) \ \ge\ 1 - e^{-\tau}$$

for all $r > r^*$. Using the definition of $g_{f_D,r}$, we thus have with probability P^n not less than $1 - e^{-\tau}$ that

$$\begin{aligned}\mathbb{E}_{\mathrm{P}} h_{\widehat{f}_D} - \mathbb{E}_{\mathrm{D}} h_{\widehat{f}_D} &< \bigl(\Upsilon(f_D) + \mathbb{E}_{\mathrm{P}} h_{\widehat{f}_D}\bigr)\left(\frac{10\varphi_n(r)}{r} + \sqrt{\frac{2V\tau}{nr^{2-\vartheta}}} + \frac{28B\tau}{3nr}\right)\\ &\quad + 10\varphi_n(r) + \sqrt{\frac{2V\tau r^\vartheta}{n}} + \frac{28B\tau}{3n}.\end{aligned}$$

By combining this estimate with (7.39) and (7.42), we then obtain that

$$\begin{aligned}\Upsilon(f_D) + \mathbb{E}_{\mathrm{P}} h_{\widehat{f}_D} &< \Upsilon(f_0) + 2\mathbb{E}_{\mathrm{P}} h_{f_0} + \left(\frac{2V\tau}{n}\right)^{\frac{1}{2-\vartheta}} + \frac{7B_0\tau}{6n} + \epsilon\\ &\quad + \bigl(\Upsilon(f_D) + \mathbb{E}_{\mathrm{P}} h_{\widehat{f}_D}\bigr)\left(\frac{10\varphi_n(r)}{r} + \sqrt{\frac{2V\tau}{nr^{2-\vartheta}}} + \frac{28B\tau}{3nr}\right)\\ &\quad + 10\varphi_n(r) + \sqrt{\frac{2V\tau r^\vartheta}{n}} + \frac{32B\tau}{3n}\end{aligned} \tag{7.44}$$

holds with probability P^n not less than $1 - 3e^{-\tau}$. Consequently, it remains to bound the various terms. To this end, we first observe that $r \geq 30\varphi_n(r)$ implies $10\varphi_n(r)r^{-1} \leq 1/3$ and $10\varphi_n(r) \leq r/3$. Moreover, $r \geq \left(\frac{72V\tau}{n}\right)^{1/(2-\vartheta)}$ yields

$$\left(\frac{2V\tau}{nr^{2-\vartheta}}\right)^{1/2} \leq \frac{1}{6} \qquad \text{and} \qquad \left(\frac{2V\tau r^{\vartheta}}{n}\right)^{1/2} \leq \frac{r}{6}\,.$$

In addition, $n \geq 72\tau$, $V \geq B^{2-\vartheta}$, and $r \geq \left(\frac{72V\tau}{n}\right)^{1/(2-\vartheta)}$ imply

$$\frac{28B\tau}{3nr} = \frac{7}{54} \cdot \frac{72\tau}{n} \cdot \frac{B}{r} \leq \frac{7}{54} \cdot \left(\frac{72\tau}{n}\right)^{\frac{1}{2-\vartheta}} \cdot \frac{V^{\frac{1}{2-\vartheta}}}{r} \leq \frac{7}{54}$$

and $\frac{32B\tau}{3n} \leq \frac{4r}{27}$. Finally, using $6 \leq 36^{1/(2-\vartheta)}$, we obtain $\left(\frac{2V\tau}{n}\right)^{1/(2-\vartheta)} \leq \frac{r}{6}$. Using these elementary estimates in (7.44), we see that

$$\Upsilon(f_D) + \mathbb{E}_{\mathrm{P}} h_{\widehat{f}_D} < \Upsilon(f_0) + 2\mathbb{E}_{\mathrm{P}} h_{f_0} + \frac{7B_0\tau}{6n} + \epsilon + \frac{17}{27}\big(\Upsilon(f_D) + \mathbb{E}_{\mathrm{P}} h_{\widehat{f}_D}\big) + \frac{22r}{27}$$

holds with probability P^n not less than $1 - 3e^{-\tau}$. We now obtain the assertion by some simple algebraic transformations together with $r > \frac{5B_0\tau}{n}$. □

Starting from Theorem 7.20, it is straightforward to recover the ERM oracle inequality (7.28) obtained in the previous section. The details are left to the interested reader.

Before we apply the general oracle inequality above to SVMs, we need the following lemma, which estimates entropy numbers for RKHSs.

Lemma 7.21 (A general entropy bound for RKHSs). *Let k be a kernel on X with RKHS H. Moreover, let $n \geq 2$, $D := (x_1, \ldots, x_n) \in X^n$, and* D *be the associated empirical measure. Then there exists a constant $K \geq 1$ independent of X, H, D, and n such that*

$$\sum_{i=1}^{\infty} 2^{i/2} e_{2^i}(\mathrm{id} : H \to L_2(\mathrm{D})) \leq K\sqrt{\ln n}\, \|k\|_{L_2(\mathrm{D})}\,.$$

Proof. We have $\operatorname{rank}(\mathrm{id} : H \to L_2(\mathrm{D})) \leq \dim L_2(\mathrm{D}) \leq n$, and hence the definition (A.29) of the approximation numbers yields $a_i(\mathrm{id} : H \to L_2(\mathrm{D})) = 0$ for all $i > n$. By Carl's inequality in the form of (A.43), we thus find that

$$\begin{aligned}\sum_{i=0}^{\infty} 2^{i/2} e_{2^i}(\mathrm{id} : H \to L_2(\mathrm{D})) &\leq c_{2,1} \sum_{i=0}^{\log_2 n} 2^{i/2} a_{2^i}(\mathrm{id} : H \to L_2(\mathrm{D})) \\ &\leq c_{2,1}\sqrt{1+\log_2 n}\left(\sum_{i=0}^{\log_2 n} 2^i a_{2^i}^2(\mathrm{id} : H \to L_2(\mathrm{D}))\right)^{\frac{1}{2}},\end{aligned}$$

where in the last step we used Hölder's inequality. Moreover, the sequence (a_i) defined by $a_i := a_i(\mathrm{id} : H \to L_2(\mathrm{D}))$, $i \geq 1$, is decreasing, and hence we have

$$\sum_{i=0}^{\log_2 n} 2^i a_{2^i}^2 \leq 2 \sum_{i=0}^{\infty} 2^{i-1} a_{2^i}^2 \leq 2 \sum_{i=1}^{\infty} \sum_{j=2^{i-1}}^{2^i-1} a_j^2 = 2 \sum_{i=1}^{\infty} a_i^2 \,.$$

Now recall that we have seen in Theorems 4.26 and 4.27 that $\mathrm{id} : H \to L_2(\mathrm{D})$ is the adjoint S_k^* of the Hilbert-Schmidt operator $S_k : L_2(\mathrm{D}) \to H$ defined by (4.17), and hence S_k^* is Hilbert-Schmidt and in particular compact. Moreover, recall (A.29) and the subsequent paragraph, where we have seen that the approximation numbers equal the singular numbers for compact operators acting between Hilbert spaces. With this information, (A.27), and the estimates above, we now find that for $K := 3c_{2,1}$ we have

$$\begin{aligned}
\sum_{i=0}^{\infty} 2^{i/2} e_{2^i}(\mathrm{id} : H \to L_2(\mathrm{D})) &\leq K \sqrt{\log_2 n} \biggl(\sum_{i=1}^{\infty} a_i^2(\mathrm{id} : H \to L_2(\mathrm{D})) \biggr)^{1/2} \\
&= K \sqrt{\log_2 n} \biggl(\sum_{i=1}^{\infty} s_i^2(S_k^* : H \to L_2(\mathrm{D})) \biggr)^{1/2} \\
&= K \sqrt{\log_2 n} \, \|S_k^*\|_{\mathsf{HS}} \,.
\end{aligned}$$

Finally, recall that an operator shares its Hilbert-Schmidt norm with its adjoint, and hence we find $\|S_k^*\|_{\mathsf{HS}} = \|S_k\|_{\mathsf{HS}} = \|k\|_{L_2(\mathrm{D})}$ by Theorem 4.27. □

Let us now formulate our first improved oracle inequality for SVMs, which holds under somewhat minimal assumptions.

Theorem 7.22 (Oracle inequality for SVMs). *Let $L : X \times Y \times \mathbb{R} \to [0, \infty)$ be a Lipschitz continuous loss with $|L|_1 \leq 1$ that can be clipped at $M > 0$ and that satisfies the supremum bound (7.35) for a constant $B \geq 1$. Furthermore, let P be a distribution on $X \times Y$ and H be a separable RKHS of a bounded measurable kernel k on X with $\|k\|_\infty \leq 1$. Then there exists a constant $K > 0$ such that for all fixed $\tau \geq 1$, $n \geq 2$, and $\lambda > 0$ the SVM associated with L and H satisfies*

$$\lambda \|f_{\mathrm{D},\lambda}\|^2 + \mathcal{R}_{L,\mathrm{P}}(\widehat{f}_{\mathrm{D},\lambda}) - \mathcal{R}_{L,\mathrm{P}}^* \leq 9A_2(\lambda) + K \frac{\ln n}{\lambda n} + \frac{15\tau}{n} \sqrt{\frac{A_2(\lambda)}{\lambda}} + \frac{300B\tau}{\sqrt{n}}$$

with probability P^n not less than $1 - 3e^{-\tau}$, where $A_2(\,\cdot\,)$ denotes the approximation error function associated with L and H.

Proof. We will use Theorem 7.20 for the regularizer $\Upsilon : H \to [0, \infty)$ defined by $\Upsilon(f) := \lambda \|f\|_H^2$. To this end, we write

$$\mathcal{F}_r := \{f \in H : \lambda \|f\|_H^2 + \mathcal{R}_{L,\mathrm{P}}(\widehat{f}\,) - \mathcal{R}_{L,\mathrm{P}}^* \leq r\} \,, \tag{7.45}$$

$$\mathcal{H}_r := \{L \circ \widehat{f} - L \circ f_{L,\mathrm{P}}^* : f \in \mathcal{F}_r\} \,. \tag{7.46}$$

From $\lambda \|f\|_H^2 \leq \lambda \|f\|_H^2 + \mathcal{R}_{L,\mathrm{P}}(\widehat{f}\,) - \mathcal{R}_{L,\mathrm{P}}^*$, we conclude $\mathcal{F}_r \subset (r/\lambda)^{1/2} B_H$, and hence we have

$$e_i(\mathcal{F}_r, L_2(\mathrm{D}_X)) \leq 2(r/\lambda)^{1/2} e_i(\mathrm{id} : H \to L_2(\mathrm{D}_X))$$

by (A.36). Consequently, Lemmas 7.17 and 7.21 yield

$$\sum_{i=0}^{\infty} 2^{\frac{i}{2}} e_{2^i}(\mathcal{H}_r, L_2(\mathrm{D})) \leq \sum_{i=0}^{\infty} 2^{\frac{i}{2}} e_{2^i}(\mathcal{F}_r, L_2(\mathrm{D}_X)) \leq K_1 \Big(\frac{r}{\lambda}\Big)^{1/2} \sqrt{\ln n}\,,$$

where $K_1 > 0$ is a universal constant. Therefore, Theorem 7.13 together with Lemma 7.14 shows that

$$\begin{aligned}
\mathrm{Rad}_D(\mathcal{H}_r, n) &\leq \sqrt{\frac{\ln 16}{n}} \Big(\sum_{i=1}^{\infty} 2^{\frac{i}{2}}\, e_{2^i}\big(\mathcal{H}_r \cup \{0\}, L_2(\mathrm{D})\big) + \sup_{h \in \mathcal{H}_r} \|h\|_{L_2(\mathrm{D})} \Big) \\
&\leq \sqrt{\frac{\ln 16}{n}} \Big(2 \sum_{i=0}^{\infty} 2^{\frac{i}{2}}\, e_{2^i}(\mathcal{H}_r, L_2(\mathrm{D})) + B \Big) \\
&\leq K_2 \Big(\frac{r}{\lambda n}\Big)^{1/2} \sqrt{\ln n} + B \sqrt{\frac{\ln 16}{n}}\,, \qquad (7.47)
\end{aligned}$$

where $K_2 > 0$ is another universal constant. Let us denote the right-hand side of (7.47) by $\varphi_n(r)$. Some elementary calculations then show that the condition $r \geq 30\varphi_n(r)$ is fulfilled for

$$r \geq \max\left\{ K_3 \frac{\ln n}{\lambda n},\ \frac{100B}{\sqrt{n}} \right\},$$

where $K_3 > 0$ is another universal constant. Moreover, the variance bound (7.36) is satisfied for $\vartheta := 0$ and $V := B^2$. In other words, the requirement $r \geq (\frac{72V\tau}{n})^{1/(2-\vartheta)}$ is fullfilled for $r \geq 9B\sqrt{\tau/n}$. Consequently, it remains to fix an $f_0 \in H$ and estimate the corresponding B_0. Let us choose $f_0 := f_{\mathrm{P},\lambda}$. Then we have

$$L(x, y, f_{\mathrm{P},\lambda}(x)) \leq L(x, y, 0) + \big| L(x, y, f_{\mathrm{P},\lambda}(x)) - L(x, y, 0) \big| \leq B + \|f_{\mathrm{P},\lambda}\|_\infty$$

for all $(x, y) \in (X \times Y)$, and combining this estimate with $\|f_{\mathrm{P},\lambda}\|_\infty \leq \|f_{\mathrm{P},\lambda}\|_H$ and

$$\lambda \|f_{\mathrm{P},\lambda}\|_H^2 \leq \lambda \|f_{\mathrm{P},\lambda}\|_H^2 + \mathcal{R}_{L,\mathrm{P}}(f_{\mathrm{P},\lambda}) - \mathcal{R}^*_{L,\mathrm{P}} = A_2(\lambda)$$

yields $\|L \circ f_{\mathrm{P},\lambda}\|_\infty \leq B + \sqrt{\lambda^{-1} A_2(\lambda)} =: B_0$. For

$$r := K_3 \frac{\ln n}{\lambda n} + \frac{100B\tau}{\sqrt{n}} + \frac{5\tau}{n} \sqrt{\frac{A_2(\lambda)}{\lambda}} + r^*\,,$$

where r^* is defined by (7.32), it is then easy to check that (7.38) is satisfied. Applying Theorem 7.20 together with $r^* \leq A_2(\lambda)$ and some elementary calculations then yields the assertion. □

It is interesting to note that Theorem 7.22 as well as the following oracle inequality also hold (in a suitably modified form) for "ϵ-approximate clipped SVMs". We refer to Exercise 7.4 for a precise statement.

Let us now show how we can improve Theorem 7.22 when additional knowledge on the RKHS in terms of entropy numbers is available.

Theorem 7.23 (Oracle inequality for SVMs using benign kernels). *Let $L : X \times Y \times \mathbb{R} \to [0,\infty)$ be a locally Lipschitz continuous loss that can be clipped at $M > 0$ and satisfies the supremum bound (7.35) for a $B > 0$. Moreover, let H be a separable RKHS of a measurable kernel over X and P be a distribution on $X \times Y$ such that the variance bound (7.36) is satisfied for constants $\vartheta \in [0,1]$, $V \geq B^{2-\vartheta}$, and all $f \in H$. Assume that for fixed $n \geq 1$ there exist constants $p \in (0,1)$ and $a \geq B$ such that*

$$\mathbb{E}_{D_X \sim \mathrm{P}_X^n} e_i(\mathrm{id} : H \to L_2(\mathrm{D}_X)) \;\leq\; a\, i^{-\frac{1}{2p}}\,, \qquad\qquad i \geq 1. \tag{7.48}$$

Finally, fix an $f_0 \in H$ and a constant $B_0 \geq B$ such that $\|L \circ f_0\|_\infty \leq B_0$. Then, for all fixed $\tau > 0$ and $\lambda > 0$, the SVM using H and L satisfies

$$\begin{aligned}\lambda\|f_{\mathrm{D},\lambda}\|_H^2 + \mathcal{R}_{L,\mathrm{P}}(\widehat{f}_{\mathrm{D},\lambda}) - \mathcal{R}^*_{L,\mathrm{P}} &\leq 9\bigl(\lambda\|f_0\|_H^2 + \mathcal{R}_{L,\mathrm{P}}(f_0) - \mathcal{R}^*_{L,\mathrm{P}}\bigr) \\ &\quad + K\Bigl(\frac{a^{2p}}{\lambda^p n}\Bigr)^{\frac{1}{2-p-\vartheta+\vartheta p}} + 3\Bigl(\frac{72V\tau}{n}\Bigr)^{\frac{1}{2-\vartheta}} + \frac{15B_0\tau}{n}\end{aligned}$$

with probability P^n not less than $1 - 3e^{-\tau}$, where $K \geq 1$ is a constant only depending on p, M, B, ϑ, and V.

Proof. We first note that it suffices to consider the case $a^{2p} \leq \lambda^p n$. Indeed, for $a^{2p} > n\lambda^p$, we have

$$\begin{aligned}\lambda\|f_{\mathrm{D},\lambda}\|_H^2 + \mathcal{R}_{L,\mathrm{P}}(\widehat{f}_{\mathrm{D},\lambda}) - \mathcal{R}^*_{L,\mathrm{P}} &\leq \lambda\|f_{\mathrm{D},\lambda}\|_H^2 + \mathcal{R}_{L,\mathrm{D}}(f_{\mathrm{D},\lambda}) + B \\ &\leq \mathcal{R}_{L,\mathrm{D}}(0) + B \\ &\leq 2B\Bigl(\frac{a^{2p}}{\lambda^p n}\Bigr)^{\frac{1}{2-p-\vartheta+\vartheta p}}\,.\end{aligned}$$

In other words, the assertion is trivially satisfied for $a^{2p} > n\lambda^p$ whenever $K \geq 2B$. Let us now define r^* by (7.32), and for $r > r^*$ we define $\mathcal{F}_r$ and $\mathcal{H}_r$ by (7.45) and (7.46), respectively. Since $\mathcal{F}_r \subset (r/\lambda)^{1/2} B_H$, Lemma 7.17 together with (7.48) and (A.36) then yields

$$\begin{aligned}\mathbb{E}_{D\sim\mathrm{P}^n} e_i(\mathcal{H}_r, L_2(\mathrm{D})) &\leq |L|_{M,1}\mathbb{E}_{D_X\sim\mathrm{P}_X^n} e_i(\mathcal{F}_r, L_2(\mathrm{D}_X)) \\ &\leq 2|L|_{M,1}\Bigl(\frac{r}{\lambda}\Bigr)^{1/2} a\, i^{-\frac{1}{2p}}\,.\end{aligned}$$

Moreover, for $f \in \mathcal{F}_r$, we have $\mathbb{E}_\mathrm{P}(L\circ\widehat{f} - L\circ f^*_{L,\mathrm{P}})^2 \leq Vr^\vartheta$, and consequently Theorem 7.16 applied to $\mathcal{H} := \mathcal{H}_r$ shows that (7.37) is satisfied for

$$\varphi_n(r) := \max\Bigg\{ C_1(p) 2^p |L|_{M,1}^p a^p \Big(\frac{r}{\lambda}\Big)^{\frac{p}{2}} (V r^{\vartheta})^{\frac{1-p}{2}} n^{-\frac{1}{2}}, \\ C_2(p) \big(2^p |L|_{M,1}^p a^p\big)^{\frac{2}{1+p}} \Big(\frac{r}{\lambda}\Big)^{\frac{p}{1+p}} B^{\frac{1-p}{1+p}} n^{-\frac{1}{1+p}} \Bigg\},$$

where $C_1(p)$ and $C_2(p)$ are the constants appearing in Theorem 7.16. Furthermore, some elementary calculations using $2 - p - \vartheta + \vartheta p \geq 1$ and $a^{2p} \leq \lambda^p n$ show that the condition $r \geq 30\varphi_n(r)$ is satisfied if

$$r \geq \tilde{K} \Big(\frac{a^{2p}}{\lambda^p n}\Big)^{\frac{1}{2-p-\vartheta+\vartheta p}},$$

where

$$\tilde{K} := \max\Big\{ \big(30 \cdot 2^p C_1(p) |L|_{M,1}^p V^{\frac{1-p}{2}}\big)^{\frac{2}{2-p-\vartheta+\vartheta p}}, 30 \cdot 120^p C_2^{1+p}(p) |L|_{M,1}^{2p} B^{1-p} \Big\}.$$

Using $r^* \leq A_2(\lambda) \leq \lambda \|f_0\|_H^2 + \mathcal{R}_{L,\mathrm{P}}(f_0) - \mathcal{R}_{L,\mathrm{P}}^*$, the assertion thus follows from Theorem 7.20 for $K := \max\{3\tilde{K}, 2B\}$. □

Let us now compare the oracle inequalities for SVMs obtained in this section with the simple oracle inequalities derived in Theorems 6.25 and 6.24. To this end, let us briefly recall the assumptions made at the end of Section 6.4, where we established the first consistency results and learning rates for SVMs: in the following, L is a Lipschitz continuous loss with $L(x, y, 0) \leq 1$ for all $(x, y) \in X \times Y$, and $|L|_1 \leq 1$. Moreover, H is a fixed separable RKHS over X having a bounded measurable kernel with $\|k\|_\infty \leq 1$. We assume that H is dense in $L_1(\mu)$ for all distributions μ on X, so that by Theorem 5.31 we have $\mathcal{R}_{L,\mathrm{P},H}^* = \mathcal{R}_{L,\mathrm{P}}^*$ for all probability measures P on $X \times Y$. We will also need the entropy number assumption

$$e_i(\mathrm{id} : H \to \ell_\infty(X)) \;\leq\; a\, i^{-\frac{1}{2p}}, \qquad\qquad i \geq 1, \tag{7.49}$$

where $a \geq 1$ and $p \in (0, 1)$ are fixed constants. Recall that by Lemma 6.21 and Exercise 6.8 this assumption is essentially equivalent to the covering number assumption (6.20). Finally, we assume in the following comparison that (λ_n) is an *a priori*[5] chosen sequence of strictly positive numbers converging to zero.

Our first goal is to consider conditions on (λ_n) ensuring universal L-risk consistency. To this end, let us first recall (see Exercise 6.9) that the oracle inequality of Theorem 6.24 ensured universal consistency if $\lambda_n^2 n \to \infty$, whereas the inequality of Theorem 6.25 ensured this for the weaker condition $\lambda_n^{1+p} n \to \infty$, where p is the exponent appearing in (7.49). Remarkably, Theorem 7.22 guarantees universal consistency for the even milder condition $\lambda_n n / \ln n \to \infty$ *without* using an entropy number assumption. Finally, if such an assumption in the form of (7.49) is satisfied, then Theorem 7.23 ensures universal consistency

[5] Data-dependent choices for λ will be discussed after the comparison.

whenever $\lambda_n^{\max\{1/2,p\}} n \to \infty$. Obviously, this is the weakest condition of the four listed. In addition, Theorem 7.23 actually allows us to replace the entropy number condition (7.49) by (7.48), which in some cases is substantially weaker, as we will see in Section 7.5.

Let us now compare the learning rates we can derive from the four oracle inequalities. To this end, we assume, as usual, that there exist constants $c > 0$ and $\beta \in (0, 1]$ such that $A_2(\lambda) \leq c\lambda^\beta$ for all $\lambda \geq 0$. Under this assumption, we have seen in Exercise 6.9 that Theorem 6.24 leads to the learning rate

$$n^{-\frac{\beta}{2\beta+2}} . \tag{7.50}$$

Moreover, at the end of Section 6.4, we derived the learning rate

$$n^{-\frac{\beta}{(2\beta+1)(1+p)}} \tag{7.51}$$

from Theorem 6.25, where p is the exponent from (7.49). On the other hand, optimizing the oracle inequality of Theorem 7.22 with the help of Lemma A.1.7 yields the learning rate

$$n^{-\frac{\beta}{\beta+1}} \ln n , \tag{7.52}$$

which is better than (7.50) by a factor of two in the exponent. Moreover, it is even better than (7.51), which was derived under more restrictive assumptions than (7.52). Finally, by assuming (7.49), or the potentially weaker condition (7.48), Theorem 7.23 applied with $f_0 := f_{\mathrm{P},\lambda}$ and $B_0 := B + \sqrt{\lambda^{-1} A_2(\lambda)}$ together with Lemma A.1.7 provides the learning rate

$$n^{-\min\{\frac{2\beta}{\beta+1}, \frac{\beta}{\beta(2-p-\vartheta+\vartheta p)+p}\}} , \tag{7.53}$$

which reduces to $n^{-\min\{\frac{2\beta}{\beta+1}, \frac{\beta}{\beta(2-p)+p}\}}$ if no variance bound assumption, i.e., $\vartheta = 0$ is made. It is simple to check that the latter is even faster than (7.52).

So far, we have only considered Lipschitz continuous losses; however, Theorem 7.23 also holds for losses that are only *locally* Lipschitz continuous. Let us now briefly describe how the rates above change when considering such losses. For the sake of simplicity we only consider the least squares loss, $Y \subset [-1, 1]$, and the choice $f_0 := f_{\mathrm{P},\lambda}$. Then, because of the growth behavior of the least squares loss, we can only choose $B_0 := 2 + 2\lambda^{-1} A_2(\lambda)$, which, by Theorem 7.23 and Example 7.3, leads to the rate

$$n^{-\min\{\beta, \frac{\beta}{\beta+p}\}} . \tag{7.54}$$

It is easy to see that the learning rates (7.52) and (7.53) are achieved for regularization sequences of the form $\lambda_n := n^{-\rho/\beta}$, where ρ is the exponent in the corresponding learning rate. To achieve these learning rates with an *a priori* chosen sequence, we therefore need to know the value of β and for (7.53) also the values of ϑ and p. Since this is unrealistic, we now demonstrate that

the TV-SVM defined in Definition 6.28 is adaptive to these parameters by choosing the regularization parameter in a data-dependent way. For brevity's sake, we focus on the situation of Theorem 7.23; however, it is straightforward to show a similar result for the more general situation of Theorem 7.22.

Theorem 7.24 (Oracle inequality for TV-SVMs and benign kernels). *Let $L : X \times Y \times \mathbb{R} \to [0,\infty)$ be a Lipschitz continuous loss with $|L|_1 \leq 1$ that can be clipped at $M > 0$ and satisfies the supremum bound (7.35) for a $B \geq 1$. Moreover, let H be a separable RKHS of a measurable kernel over X with $\|k\|_\infty \leq 1$ and let P be a distribution on $X \times Y$ such that the variance bound (7.36) is satisfied for constants $\vartheta \in [0,1]$, $V \geq B^{2-\vartheta}$, and all $f \in H$. We further assume that H is dense in $L_1(\mathrm{P}_X)$. In addition, let $p \in (0,1)$ and $a \geq B$ be constants such that*

$$\mathbb{E}_{D_X \sim \mathrm{P}_X^n} e_i(\mathrm{id} : H \to L_2(\mathrm{D}_X)) \;\leq\; a\, i^{-\frac{1}{2p}}\,, \qquad n \geq 1,\, i \geq 1.$$

Moreover, for $n \geq 4$ and $\varepsilon > 0$, let $\Lambda_n \subset (0,1]$ be a finite ε-net of $(0,1]$ with cardinality $|\Lambda_n|$. For fixed $\tau \geq 1$, we then have for every corresponding, measurable TV-SVM with probability P^n not less than, $1 - e^{-\tau}$ that

$$\begin{aligned}\mathcal{R}_{L,\mathrm{P}}(\widehat{f}_{\mathrm{D}_1,\lambda_{D_2}}) - \mathcal{R}^*_{L,\mathrm{P}} < 6 \inf_{\lambda \in (0,1]} \biggl(9A_2(\lambda) + K\Bigl(\frac{2a^{2p}}{\lambda^p n}\Bigr)^{\frac{1}{2-p-\vartheta+\vartheta p}} + \frac{30\tau_n}{n}\sqrt{\frac{A_2(\lambda)}{\lambda}} \biggr) \\ + 6A_2(2\varepsilon) + 45\Bigl(\frac{64V\tau_n}{n}\Bigr)^{\frac{1}{2-\vartheta}} + \frac{180B\tau_n}{n},\end{aligned}$$

where D_1 and D_2 are the training and validation set built from $D \in (X \times Y)^n$, K is the constant appearing in Theorem 7.23, and $\tau_n := \tau + \ln(1 + 3|\Lambda_n|)$.

Consequently, if we use ε_n-nets Λ_n with $\varepsilon_n \to 0$ and $n^{-1}\ln(|\Lambda_n|) \to 0$, the resulting TV-SVM is consistent. Moreover, for $\varepsilon_n \leq 1/n^2$ and $|\Lambda_n|$ growing polynomially in n, the TV-SVM learns with rate (7.53) if there exist constants $c > 0$ and $\beta \in (0,1]$ such that $A_2(\lambda) \leq c\lambda^\beta$ for all $\lambda \geq 0$.

Note that it is easy to derive modifications of Theorem 7.24 that consider locally Lipschitz continuous loss functions. For example, for the least squares loss we only have to replace the term $\sqrt{\lambda^{-1}A_2(\lambda)}$ by $2\lambda^{-1}A_2(\lambda)$ in the oracle inequality. Moreover, for this loss the conditions ensuring consistency remain unchanged and the resulting learning rate becomes (7.54).

Proof. For $f_0 := f_{\mathrm{P},\lambda}$, we have already seen in the proof of Theorem 7.22 that $\|L \circ f_0\|_\infty \leq B + \sqrt{\lambda^{-1}A_2(\lambda)}$. Since $m := \lfloor n/2 \rfloor + 1$ implies $m \geq n/2$, we hence see by Theorem 7.23 that with probability P^m not less than $1 - 3|\Lambda_n|e^{-\tau}$ we have

$$\begin{aligned}\mathcal{R}_{L,\mathrm{P}}(\widehat{f}_{\mathrm{D}_1,\lambda}) - \mathcal{R}^*_{L,\mathrm{P}} \leq 9A_2(\lambda) + K\Bigl(\frac{2a^{2p}}{\lambda^p n}\Bigr)^{\frac{1}{2-p-\vartheta+\vartheta p}} + \frac{30\tau}{n}\sqrt{\frac{A_2(\lambda)}{\lambda}} \\ + 3\Bigl(\frac{144V\tau}{n}\Bigr)^{\frac{1}{2-\vartheta}} + \frac{30B\tau}{n}\end{aligned}$$

for all $\lambda \in \Lambda_n$ *simultaneously*. Moreover, $n \geq 4$ implies $n - m \geq n/2 - 1 \geq n/4$, and therefore Theorem 7.2 shows that, for *fixed* $D_1 \in (X \times Y)^m$ and $\tilde{\tau}_n := \tau + \ln(1 + |\Lambda_n|)$, the probability P^{n-m} of having a $D_2 \in (X \times Y)^{n-m}$ such that

$$\mathcal{R}_{L,\mathrm{P}}(\widehat{f}_{\mathrm{D}_1,\lambda_{D_2}}) - \mathcal{R}^*_{L,\mathrm{P}} \leq 6 \inf_{\lambda \in \Lambda_n} \bigl(\mathcal{R}_{L,\mathrm{P}}(\widehat{f}_{\mathrm{D}_1,\lambda}) - \mathcal{R}^*_{L,\mathrm{P}}\bigr) + 4\biggl(\frac{32V\tilde{\tau}_n}{n}\biggr)^{\frac{1}{2-\vartheta}}$$

is not less than $1 - e^{-\tau}$. Combining both estimates, we conclude that with probability P^n not less than $(1 - 3|\Lambda_n|e^{-\tau})(1 - e^{-\tau})$ we have

$$\begin{aligned}\mathcal{R}_{L,\mathrm{P}}(\widehat{f}_{\mathrm{D}_1,\lambda_{D_2}}) - \mathcal{R}^*_{L,\mathrm{P}} &\leq 6 \inf_{\lambda \in \Lambda_n} \biggl(9A_2(\lambda) + K\Bigl(\frac{2a^{2p}}{\lambda^p n}\Bigr)^{\frac{1}{2-p-\vartheta+\vartheta p}} + \frac{30\tau}{n}\sqrt{\frac{A_2(\lambda)}{\lambda}}\biggr) \\ &\quad + 41\Bigl(\frac{64V\tau}{n}\Bigr)^{\frac{1}{2-\vartheta}} + \frac{180B\tau}{n} + 4\biggl(\frac{32V\tilde{\tau}_n}{n}\biggr)^{\frac{1}{2-\vartheta}}.\end{aligned}$$

Now recall that $\lambda \mapsto \lambda^{-1}A_2(\lambda)$ is decreasing by Lemma 5.15, and hence an almost literal repetition of the proof of Lemma 6.30 yields

$$\begin{aligned}&\inf_{\lambda \in \Lambda_n} \biggl(9A_2(\lambda) + K\Bigl(\frac{2a^{2p}}{\lambda^p n}\Bigr)^{\frac{1}{2-p-\vartheta+\vartheta p}} + \frac{30\tau}{n}\sqrt{\frac{A_2(\lambda)}{\lambda}}\biggr) \\ &\leq A_2(2\varepsilon) + \inf_{\lambda \in (0,1]} \biggl(9A_2(\lambda) + K\Bigl(\frac{2a^{2p}}{\lambda^p n}\Bigr)^{\frac{1}{2-p-\vartheta+\vartheta p}} + \frac{30\tau}{n}\sqrt{\frac{A_2(\lambda)}{\lambda}}\biggr).\end{aligned}$$

Using $(1 - 3|\Lambda_n|e^{-\tau})(1 - e^{-\tau}) \geq 1 - (1 + 3|\Lambda_n|)e^{-\tau}$ together with a variable transformation that adjusts the probability P^n to be not less than $1 - e^{-\tau}$ and some simple estimates then yields the asserted oracle inequality. The other assertions are direct consequences of this oracle inequality. □

7.5 Some Bounds on Average Entropy Numbers

The goal of this section is to illustrate that the entropy assumption (7.48) used in Theorems 7.23 and 7.24 is weaker than the entropy bound (7.49) used in the simple analysis of Section 6.4. To this end, we first relate the average entropy numbers in (7.48) to the eigenvalues of the integral operator defined by the kernel of the RKHS used.

Let us begin by establishing some preparatory lemmas. To this end, we fix a Hilbert space H. For $v, w \in H$, we define $v \otimes w : H \to H$ by $v \otimes w(u) := \langle u, v\rangle w$, $u \in H$. Obviously, $v \otimes w$ is bounded and linear with $\operatorname{rank} v \otimes w \leq 1$, and hence it is a Hilbert-Schmidt operator, i.e., $v \otimes w \in \mathsf{HS}(H)$. Since in the following we need various facts on Hilbert-Schmidt operators, we encourage the reader to review the corresponding material presented at the end of Section A.5.2.

Lemma 7.25 (Feature maps of squared kernels). *Let H be an RKHS over X with kernel k and canonical feature map $\Phi : X \to H$. Then $\Psi : X \to \mathsf{HS}(H)$ defined by $\Psi(x) := \Phi(x) \otimes \Phi(x)$, $x \in X$, is a feature map of the squared kernel $k^2 : X \times X \to \mathbb{R}$, and we have $\|\Psi(x)\|_{\mathsf{HS}} = \|\Phi(x)\|_H^2$ for all $x \in X$.*

Proof. For $f \in H$ and $x \in X$, the reproducing property yields

$$\Psi(x)f := \Psi(x)(f) = \langle f, \Phi(x)\rangle_H \, \Phi(x) = f(x)\Phi(x)\,. \tag{7.55}$$

Let $(e_i)_{i\in I}$ be an ONB of H. By the definition (A.28) of the inner product of $\mathsf{HS}(H)$ and (4.9), we then obtain

$$\begin{aligned}\langle \Psi(x), \Psi(x')\rangle_{\mathsf{HS}(H)} = \sum_{i\in I} \langle \Psi(x)e_i, \Psi(x')e_i\rangle_H &= \sum_{i\in I} \langle e_i(x)\Phi(x), e_i(x')\Phi(x')\rangle_H \\ &= \langle \Phi(x), \Phi(x')\rangle_H \sum_{i\in I} e_i(x)e_i(x') \\ &= k^2(x, x') \end{aligned} \tag{7.56}$$

for all $x, x' \in X$, i.e., we have shown the first assertion. The second assertion follows from (7.56) and $\|\Psi(x)\|_{\mathsf{HS}} = \sqrt{k^2(x,x)} = k(x,x) = \|\Phi(x)\|_H^2$. □

Let us now assume that k is a measurable kernel on X with separable RKHS H and μ is a probability measure on X with $\|k\|_{L_2(\mu)} < \infty$. In the following, we write $S_{k,\mu} : L_2(\mu) \to H$ for the integral operator defined in (4.17) in order to emphasize that this operator depends not only on k but also on μ. Analogously, we write $T_{k,\mu} := S_{k,\mu}^* \circ S_{k,\mu} : L_2(\mu) \to L_2(\mu)$ for the integral operator considered in Theorem 4.27. Finally, we also need the **covariance operator** $C_{k,\mu} := S_{k,\mu} \circ S_{k,\mu}^* : H \to H$. The next lemma relates $C_{k,\mu}$ to the feature map Ψ of k^2.

Lemma 7.26. *Let k be a measurable kernel on X with separable RKHS H and ν be a probability measure on X such that $\|k\|_{L_2(\nu)} < \infty$. Then $\Psi : X \to \mathsf{HS}(H)$ defined in Lemma 7.25 is Bochner ν-integrable and $\mathbb{E}_\nu \Psi = C_{k,\nu}$.*

Proof. Obviously, Ψ is measurable, and since (7.56) implies

$$\mathbb{E}_{x\sim\nu} \|\Psi(x)\|_{\mathsf{HS}} = \mathbb{E}_{x\sim\nu} k(x,x) = \|k\|_{L_2(\nu)} < \infty\,,$$

we see that Ψ is Bochner integrable with respect to ν. Let us now fix an $f \in H$. Since by Theorem 4.26 the operator $S_{k,\nu}^*$ equals $\mathrm{id} : H \to L_2(\nu)$, we then have $C_{k,\nu} f = S_{k,\nu} \circ S_{k,\nu}^* f = \mathbb{E}_\nu f\Phi$. In addition, considering the bounded linear operator $\delta_f : \mathsf{HS}(H) \to H$ defined by $\delta_f(S) := Sf$, $S \in \mathsf{HS}(H)$, we find by (A.32) and (7.55) that

$$(\mathbb{E}_\nu \Psi)(f) = \delta_f\big(\mathbb{E}_{x\sim\nu}\Psi(x)\big) = \mathbb{E}_{x\sim\nu}\delta_f(\Psi(x)) = \mathbb{E}_{x\sim\nu}\big(\Psi(x)(f)\big) = \mathbb{E}_\nu f\Phi\,.$$

By combining our considerations, we then obtain $\mathbb{E}_\nu \Psi = C_{k,\nu}$. □

Before we can state the first main result of this section, we finally need the following two technical lemmas.

Lemma 7.27. *Let $(\alpha_i) \subset [0,1]$ be a sequence such that $\sum_{i=1}^{\infty} \alpha_i = m$ for some $m \in \mathbb{N}$ and $(\lambda_i) \subset [0,\infty)$ be a decreasing sequence. Then we have*

$$\sum_{i=1}^{\infty} \alpha_i \lambda_i \leq \sum_{i=1}^{m} \lambda_i \,.$$

Proof. Let us define $\gamma_i := \lambda_i - \lambda_m$, $i \geq 1$. Then we have $\gamma_i \geq 0$ if $i \leq m$ and $\gamma_i \leq 0$ if $i \geq m$, and consequently we obtain

$$\sum_{i=1}^{\infty} \alpha_i \gamma_i \leq \sum_{i=1}^{m} \alpha_i \gamma_i \leq \sum_{i=1}^{m} \gamma_i \,.$$

Moreover, we have

$$\sum_{i=1}^{\infty} \alpha_i \lambda_i - m\lambda_m = \sum_{i=1}^{\infty} \alpha_i \lambda_i - \sum_{i=1}^{\infty} \alpha_i \lambda_m = \sum_{i=1}^{\infty} \alpha_i \gamma_i$$

and

$$\sum_{i=1}^{m} \gamma_i = \sum_{i=1}^{m} \lambda_i - \sum_{i=1}^{m} \lambda_m = \sum_{i=1}^{m} \lambda_i - m\lambda_m \,,$$

and by combining all formulas, we then obtain the assertion. □

Given a compact and self-adjoint operator $T \in \mathcal{L}(H)$, we now investigate the set of non-zero eigenvalues $\{\lambda_i(T) : i \in I\}$ considered in the Spectral Theorem A.5.13. Following the notation of this theorem, we assume in the following that either $I = \{1, 2, \ldots, n\}$ or $I = \mathbb{N}$. In order to unify our considerations, we further need the **extended sequence of eigenvalues** of T, which in the case of finite I is the sequence $\lambda_1(T), \lambda_2(T), \ldots, \lambda_{|I|}(T), 0, \ldots$, and in the case of infinite I is the sequence $(\lambda_i(T))_{i \geq 1}$. With these notations, we can now establish the last preparatory lemma.

Lemma 7.28. *Let $T : H \to H$ be a compact, positive, and self-adjoint operator on a separable Hilbert space H and $(\lambda_i(T))_{i \geq 1}$ be its extended sequence of eigenvalues. Then, for all $m \geq 1$ and $\bar{m} := \min\{m, \dim H\}$, we have*

$$\sum_{i=1}^{m} \lambda_i(T) \;=\; \max_{\substack{V \text{subspace of } H \\ \dim V = \bar{m}}} \langle P_V, T \rangle_{\mathsf{HS}} \,.$$

Proof. Let $(e_i)_{i \in J}$ be an ONB of H with $I \subset J$ such that the subfamily $(e_i)_{i \in I}$ is the ONS provided by Theorem A.5.13. In addition, we assume for notational convenience that $J = \{1, \ldots, \dim H\}$ if $\dim H < \infty$ and $J = \mathbb{N}$ if $\dim H = \infty$ and $|I| < \infty$. Note that in both cases we have $\lambda_i(T) = 0$ for all

$i \in J\backslash I$ since we consider extended sequences of eigenvalues. Moreover, in the case $\dim H = |I| = \infty$, we define $\lambda_i(T) := 0$ for $i \in J \setminus I$. Let us now write $V_0 := \operatorname{span}\{e_1, \ldots, e_{\bar{m}}\}$, where we note that our notational assumptions ensure $\{1, \ldots, \bar{m}\} \subset J$. Obviously, the orthogonal projection $P_{V_0} : H \to H$ onto V_0 satisfies $P_{V_0} e_i = e_i$ if $i \in \{1, \ldots, \bar{m}\}$ and $P_{V_0} e_i = 0$ otherwise. Moreover, we have $Te_i = \lambda_i(T) e_i$ for all $i \in I$, and the spectral representation (A.23) shows $Te_i = 0$ for all $i \in J\backslash I$. Since our notational assumptions ensured $\lambda_i(T) = 0$ for all $i \in J \setminus I$, we thus find $Te_i = \lambda_i(T) e_i$ for all $i \in J$. The definition (A.28) of the inner product in $\mathsf{HS}(H)$ together with $\lambda_i(T) = 0$ for $i \in \mathbb{N}\backslash\{1, \ldots, \dim H\}$ hence yields

$$\sum_{i=1}^{m} \lambda_i(T) = \sum_{i=1}^{\bar{m}} \lambda_i(T) = \sum_{i \in J} \langle P_{V_0} e_i, Te_i \rangle_H = \langle P_{V_0}, T \rangle_{\mathsf{HS}} \leq \max_{\substack{V \text{subspace of } H \\ \dim V = \bar{m}}} \langle P_V, T \rangle_{\mathsf{HS}} \,.$$

Conversely, if V is an arbitrary subspace of H with $\dim V = \bar{m}$, the self-adjointness of the orthogonal projection P_V onto V shows that for $\alpha_i := \langle P_V e_i, e_i \rangle_H = \langle P_V^2 e_i, e_i \rangle_H$ we have

$$\alpha_i = \langle P_V^2 e_i, e_i \rangle_H = \langle P_V^* \circ P_V e_i, e_i \rangle_H = \langle P_V e_i, P_V e_i \rangle_H = \|P_V e_i\|_H^2 \in [0,1]$$

for all $i \in J$. Let us write $\alpha_i := 0$ if $\dim H < \infty$ and $i > \dim H$. Then we observe

$$\sum_{i=1}^{\infty} \alpha_i = \sum_{i \in J} \|P_V e_i\|_H^2 = \|P_V\|_{\mathsf{HS}}^2 = \bar{m} \,.$$

From this, Lemma 7.27, and $Te_i = \lambda_i(T) e_i$ for all $i \in J$, we conclude that

$$\sum_{i=1}^{m} \lambda_i(T) = \sum_{i=1}^{\bar{m}} \lambda_i(T) \geq \sum_{i=1}^{\infty} \alpha_i \lambda_i(T) = \sum_{i \in J} \langle P_V e_i, Te_i \rangle_H = \langle P_V, T \rangle_{\mathsf{HS}} \,,$$

and hence we have shown the assertion. □

Let us now formulate our first main result of this section, which relates the average eigenvalues of the empirical integral operators $T_{k,\mathrm{D}}$, $D \in X^n$, with the eigenvalues of the infinite-sample integral operator $T_{k,\mu}$.

Theorem 7.29 (Eigenvalues of empirical integral operators). *Let k be a measurable kernel on X with separable RKHS H and μ be a probability measure on X such that $\|k\|_{L_2(\mu)} < \infty$. Then we have*

$$\mathbb{E}_{D \sim \mu^n} \sum_{i=1}^{\infty} \lambda_i(T_{k,\mathrm{D}}) = \sum_{i=1}^{\infty} \lambda_i(T_{k,\mu}) < \infty \tag{7.57}$$

for the extended sequences of eigenvalues. Furthermore, for all $m \geq 1$, these extended sequences of eigenvalues satisfy

$$\mathbb{E}_{D\sim\mu^n}\sum_{i=m}^{\infty}\lambda_i(T_{k,\mathrm{D}}) \le \sum_{i=m}^{\infty}\lambda_i(T_{k,\mu}) \tag{7.58}$$

and

$$\mathbb{E}_{D\sim\mu^n}\sum_{i=1}^{m}\lambda_i(T_{k,\mathrm{D}}) \ge \sum_{i=1}^{m}\lambda_i(T_{k,\mu})\,. \tag{7.59}$$

Proof. Let us begin with some preliminary considerations. To this end, let $(e_j)_{j\in J}$ be an ONB of H and ν be an arbitrary probability measure on X with $\|k\|_{L_2(\nu)} < \infty$. Recall that we have seen in front of the Spectral Theorem A.5.13 that the integral operator $T_{k,\nu} = S^*_{k,\nu}\circ S_{k,\nu}$ and the covariance operator $C_{k,\nu} = S_{k,\nu}\circ S^*_{k,\nu}$ have exactly the same non-zero eigenvalues with the same geometric multiplicities. Consequently, their extended sequences of eigenvalues coincide. Moreover, by (A.25), (A.27), (A.26), and Theorem 4.26, we find

$$\sum_{i=1}^{\infty}\lambda_i(T_{k,\nu}) = \sum_{i=1}^{\infty}s_i^2(S^*_{k,\nu}) = \|S^*_{k,\nu}\|^2_{\mathsf{HS}} = \sum_{j\in J}\|S^*_{k,\nu}e_j\|^2_{L_2(\nu)} = \sum_{j\in J}\mathbb{E}_\nu e_j^2\,,$$

where we note that $\|S^*_{k,\nu}\|_{\mathsf{HS}} = \|S_{k,\nu}\|_{\mathsf{HS}} < \infty$ by Theorem 4.27. Applying the equality above twice, we now obtain

$$\mathbb{E}_{D\sim\mu^n}\sum_{i=1}^{\infty}\lambda_i(T_{k,\mathrm{D}}) = \sum_{j\in J}\mathbb{E}_{D\sim\mu^n}\mathbb{E}_{\mathrm{D}}e_j^2 = \sum_{j\in J}\mathbb{E}_\mu e_j^2 = \sum_{i=1}^{\infty}\lambda_i(T_{k,\mu}),$$

i.e., we have shown (7.57). In order to prove (7.59), we first observe that Lemma 7.26 together with Lemma 7.28 and $\lambda_i(T_{k,\nu}) = \lambda_i(C_{k,\nu})$ for all $i \ge 1$ implies

$$\sum_{i=1}^{m}\lambda_i(T_{k,\nu}) = \max_{\substack{V\text{subspace of } H\\ \dim V=\bar m}}\langle P_V, C_{k,\nu}\rangle_{\mathsf{HS}} = \max_{\substack{V\text{subspace of } H\\ \dim V=\bar m}}\mathbb{E}_\nu\langle P_V, \Psi\rangle_{\mathsf{HS}}\,,$$

where again ν is an arbitrary probability measure on X with $\|k\|_{L_2(\nu)} < \infty$ and $\bar m := \min\{m, \dim H\}$. From this we conclude that

$$\begin{aligned}
\mathbb{E}_{D\sim\mu^n}\sum_{i=1}^{m}\lambda_i(T_{k,\mathrm{D}}) &= \mathbb{E}_{D\sim\mu^n}\max_{\substack{V\text{subspace of } H\\ \dim V=\bar m}}\mathbb{E}_{\mathrm{D}}\langle P_V, \Psi\rangle_{\mathsf{HS}}\\
&\ge \max_{\substack{V\text{subspace of } H\\ \dim V=\bar m}}\mathbb{E}_{D\sim\mu^n}\mathbb{E}_{\mathrm{D}}\langle P_V, \Psi\rangle_{\mathsf{HS}}\\
&= \max_{\substack{V\text{subspace of } H\\ \dim V=\bar m}}\mathbb{E}_{\mu}\langle P_V, \Psi\rangle_{\mathsf{HS}}\\
&= \sum_{i=1}^{m}\lambda_i(T_{k,\mu})\,.
\end{aligned}$$

Finally, (7.58) is a direct consequence of (7.57) and (7.59). □

Let us now recall that we have seen in Sections A.5.2 and A.5.6 that there is an intimate relationship between eigenvalues, singular numbers, approximation numbers, and entropy numbers. This relationship together with the preceding theorem leads to the second main result of this section.

Theorem 7.30 (Average entropy numbers of RKHSs). *Let k be a measurable kernel on X with separable RKHS H and μ be a probability measure on X such that $\|k\|_{L_2(\mu)} < \infty$. Then for all $0 < p < 1$ there exists a constant $c_p \geq 1$ only depending on p such that for all $n \geq 1$ and $m \geq 1$ we have*

$$\mathbb{E}_{D\sim\mu^n} e_m(S^*_{k,\mathrm{D}}) \leq c_p\, m^{-1/p} \sum_{i=1}^{\min\{m,n\}} i^{1/p-1} \left(\frac{1}{i} \sum_{j=i}^{\infty} e_j^2(S^*_{k,\mu}) \right)^{1/2}.$$

Proof. For $q := 1$, Carl's inequality (A.42) shows that there exists a constant $c_p > 0$ such that for $m, n \geq 1$ and all $D \in X^n$ we have

$$\sum_{i=1}^{m} i^{1/p-1} e_i(S^*_{k,\mathrm{D}}) \leq c_p \sum_{i=1}^{m} i^{1/p-1} a_i(S^*_{k,\mathrm{D}}) = c_p \sum_{i=1}^{\min\{m,n\}} i^{1/p-1} a_i(S^*_{k,\mathrm{D}}),$$

where in the last step we used that $n \geq \operatorname{rank} S^*_{k,D}$ implies $a_i(S^*_{k,\mathrm{D}}) = 0$ for all $i > n$. Moreover, for $M := \min\{m, n\}$ and $\tilde{M} := \lfloor (M+1)/2 \rfloor$, we have

$$\begin{aligned}\sum_{i=1}^{M} i^{1/p-1} a_i(S^*_{k,\mathrm{D}}) &\leq \sum_{i=1}^{\tilde{M}} (2i-1)^{1/p-1} a_{2i-1}(S^*_{k,\mathrm{D}}) + \sum_{i=1}^{\tilde{M}} (2i)^{1/p-1} a_{2i}(S^*_{k,\mathrm{D}}) \\ &\leq 2^{1/p} \sum_{i=1}^{M} i^{1/p-1} a_{2i-1}(S^*_{k,\mathrm{D}}).\end{aligned}$$

Now recall from the end of Section A.5.2 that $a_i^2(S^*_{k,\mathrm{D}}) = s_i^2(S^*_{k,\mathrm{D}}) = s_i(S^*_{k,\mathrm{D}} S_{k,\mathrm{D}}) = \lambda_i(T_{k,\mathrm{D}})$ for all $i \geq 1$ and $D \in X^n$, and hence we obtain

$$\begin{aligned}\sum_{i=1}^{m} i^{1/p-1} \mathbb{E}_{D\sim\mu^n} e_i(S^*_{k,\mathrm{D}}) &\leq 2^{1/p} c_p \sum_{i=1}^{M} i^{1/p-1} \mathbb{E}_{D\sim\mu^n} a_{2i-1}(S^*_{k,\mathrm{D}}) \\ &\leq 2^{1/p} c_p \sum_{i=1}^{M} i^{1/p-1} \big(\mathbb{E}_{D\sim\mu^n} \lambda_{2i-1}(T_{k,\mathrm{D}}) \big)^{1/2}.\end{aligned}$$

Now for each $D \in X^n$ the sequence $(\lambda_i(T_{k,\mathrm{D}}))_{i\geq 1}$ is monotonically decreasing and hence so is $(\mathbb{E}_{D\sim\mu^n} \lambda_i(T_{k,\mathrm{D}}))_{i\geq 1}$. By Theorem 7.29, we hence find

$$i\, \mathbb{E}_{D\sim\mu^n} \lambda_{2i-1}(T_{k,\mathrm{D}}) \leq \sum_{j=i}^{2i-1} \mathbb{E}_{D\sim\mu^n} \lambda_j(T_{k,\mathrm{D}}) \leq \sum_{j=i}^{\infty} \lambda_j(T_{k,\mu})$$

for all $i \geq 1$, and consequently we obtain

$$\sum_{i=1}^{M} i^{1/p-1} \bigl(\mathbb{E}_{D\sim\mu^n} \lambda_{2i-1}(T_{k,\mathrm{D}})\bigr)^{1/2} \le \sum_{i=1}^{M} i^{1/p-1} \left(\frac{1}{i} \sum_{j=i}^{\infty} \lambda_j(T_{k,\mu})\right)^{1/2}.$$

Moreover, we have $\lambda_j(T_{k,\mu}) = s_i^2(S_{k,\mu}^*) = a_j^2(S_{k,\mu}^*) \le 4e_j^2(S_{k,\mu}^*)$, where in the last step we used (A.44). Combining the estimates above, we hence obtain

$$\sum_{i=1}^{m} i^{1/p-1} \mathbb{E}_{D\sim\mu^n} e_i(S_{k,\mathrm{D}}^*) \le 2^{1/p+1} c_p \sum_{i=1}^{M} i^{1/p-1} \left(\frac{1}{i} \sum_{j=i}^{\infty} e_j^2(S_{k,\mu}^*)\right)^{1/2}.$$

Finally, for $\tilde{m} := \lfloor (m+1)/2 \rfloor$, the monotonicity of the entropy numbers yields

$$\tilde{m}^{\frac{1}{p}} \, \mathbb{E}_{D\sim\mu^n} e_m(S_{k,\mathrm{D}}^*) \le \sum_{i=\tilde{m}}^{m} i^{\frac{1}{p}-1} \, \mathbb{E}_{D\sim\mu^n} e_i(S_{k,\mathrm{D}}^*) \le \sum_{i=1}^{m} i^{\frac{1}{p}-1} \, \mathbb{E}_{D\sim\mu^n} e_i(S_{k,\mathrm{D}}^*),$$

and since $m/2 \le \lfloor (m+1)/2 \rfloor = \tilde{m}$, we hence obtain the assertion. □

With the help of Theorem 7.30, we can now formulate the following condition, which ensures the average entropy number assumption (7.48).

Corollary 7.31. *Let k be a measurable kernel on X with separable RKHS H and μ be a probability measure on X such that $\|k\|_{L_2(\mu)} < \infty$. Assume that there exist constants $0 < p < 1$ and $a \ge 1$ such that*

$$e_i(\mathrm{id} : H \to L_2(\mu)) \le a \, i^{-\frac{1}{2p}}, \qquad i \ge 1. \tag{7.60}$$

Then there exists a constant $c_p > 0$ only depending on p such that

$$\mathbb{E}_{D\sim\mu^n} e_i(\mathrm{id} : H \to L_2(\mathrm{D})) \le c_p \, a \, \bigl(\min\{i,n\}\bigr)^{\frac{1}{2p}} \, i^{-\frac{1}{p}}, \qquad i,n \ge 1.$$

Proof. By Theorem 4.26, the operator $S_{k,\nu}^*$ coincides with $\mathrm{id} : H \to L_2(\nu)$ for all distributions ν with $\|k\|_{L_2(\mu)} < \infty$. Since $0 < p < 1$, it is then easy to see that there exists a constant $\tilde{c}_p$ such that

$$\frac{1}{i} \sum_{j=i}^{\infty} e_j^2(S_{k,\mu}^*) \le a^2 \cdot \frac{1}{i} \sum_{j=i}^{\infty} j^{-\frac{1}{p}} \le \tilde{c}_p^2 \, a^2 \, i^{-\frac{1}{p}}, \qquad i \ge 1.$$

Using Theorem 7.30 and $\frac{1}{2p} - 1 > -1$, we hence find a constant $c_p' > 0$ such that

$$\mathbb{E}_{D\sim\mu^n} e_m(S_{k,\mathrm{D}}^*) \le c_p \tilde{c}_p \, a \, m^{-\frac{1}{p}} \sum_{i=1}^{\min\{m,n\}} i^{\frac{1}{2p}-1} \le c_p' \, a \, \bigl(\min\{m,n\}\bigr)^{\frac{1}{2p}} \, m^{-\frac{1}{p}}. \quad \square$$

Considering the proofs of Theorem 7.30 and Corollary 7.31, it is not hard to see that we can replace the entropy bound (7.60) by a bound on the *eigenvalues* of $T_{K,\mu}$. We refer to Exercise 7.7 for details.

Let us now return to the main goal of this section, which is to illustrate that the average entropy assumption (7.48) is weaker than the uniform entropy bound (7.49). To this end, we need the following definition.

Definition 7.32. *We say that a distribution μ on $\mathbb{R}^d$ has* ***tail exponent*** $\tau \in [0,\infty]$ *if*

$$\mu\bigl(\mathbb{R}^d \backslash r B_{\ell_2^d}\bigr) \;\le\; r^{-\tau}\,, \qquad\qquad r > 0. \tag{7.61}$$

Obviously, every distribution has tail exponent $\tau = 0$. On the other hand, a distribution has tail exponent $\tau = \infty$ if and only if the support of μ is contained in the closed unit ball $B_{\ell_2^d}$. Finally, the intermediate case $\tau \in (0,\infty)$ describes how much mass μ has *outside* the scaled Euclidean balls $rB_{\ell_2^d}$, $r > 1$.

Before we return to entropy numbers, we need another notation. To this end, let μ be a distribution on $\mathbb{R}^d$ and $X \subset \mathbb{R}^d$ be a measurable set with $\mu(X) > 0$. Then we define the distribution μ_X on $\mathbb{R}^d$ by

$$\mu_X(A) := \mu(A \cap X)/\mu(X)\,, \qquad\qquad A \subset \mathbb{R}^d \text{ measurable.}$$

With this notation, we can now formulate the following result.

Theorem 7.33 (Entropy numbers on unbounded domains). *Let H be an RKHS of a bounded kernel k on $\mathbb{R}^d$ with $\|k\|_\infty = 1$ and μ be a distribution on $\mathbb{R}^d$ that has tail exponent $\tau \in (0,\infty]$. We write $B := B_{\ell_2^d}$ and assume that there exist constants $a \ge 1$, $c \ge 1$, $\varsigma > 0$, and $p \in (0,1)$ such that*

$$e_i(\mathrm{id} : H \to L_2(\mu_{rB})) \;\le\; c\, a^{\varsigma}\, r^{\varsigma} i^{-\frac{1}{2p}}\,, \qquad\qquad i \ge 1,\, r \ge a^{-1}. \tag{7.62}$$

Then there exists a constant $c_{\varsigma,\tau} \ge 1$ only depending on ς and τ such that

$$e_i(\mathrm{id} : H \to L_2(\mu)) \;\le\; c\, c_{\varsigma,\tau}\, a^{\frac{\varsigma\tau}{2\varsigma+\tau}}\, i^{-\frac{2p\varsigma+\tau}{4p\varsigma+2p\tau}}\,, \qquad\qquad i \ge 1\,.$$

Proof. Obviously, for $\tau = \infty$, there is nothing to prove, and hence we assume $\tau < \infty$. Let us now fix an $\varepsilon > 0$, an $r \ge a^{-1}$, and an integer $i \ge 1$. Moreover, let $f_1, \ldots, f_{2^{i-1}}$ be an $(1+\varepsilon)e_i(B_H, L_2(\mu_{rB}))$-net of B_H and $f'_1, \ldots, f'_{2^{i-1}}$ be an $(1+\varepsilon)e_i(B_H, L_2(\mu_{\mathbb{R}^d\backslash rB}))$-net of B_H. For $j, l \in \{1, \ldots, 2^{i-1}\}$, we write

$$f_j \diamond f'_l := \mathbf{1}_{rB} f_j + \mathbf{1}_{\mathbb{R}^d\backslash rB} f'_l\,,$$

i.e., $f_j \diamond f'_l$ equals f_j on the scaled Euclidean ball rB, while it equals f'_l on the complement of this ball. Let us now investigate how well these functions approximate B_H in $L_2(\mu)$. To this end, we fix a $g \in B_H$. Then there exist two indexes $j, l \in \{1, \ldots, 2^{i-1}\}$ such that

$$\begin{aligned} \|g - f_j\|_{L_2(\mu_{rB})} &\le (1+\varepsilon) e_i(B_H, L_2(\mu_{rB}))\,, \\ \|g - f'_l\|_{L_2(\mu_{\mathbb{R}^d\backslash rB})} &\le (1+\varepsilon) e_i(B_H, L_2(\mu_{\mathbb{R}^d\backslash rB}))\,. \end{aligned}$$

With these estimates, we obtain

$$\begin{aligned} \|g - f_j \diamond f'_l\|^2_{L_2(\mu)} &= \mu(rB)\|g - f_j\|^2_{L_2(\mu_{rB})} + \mu(\mathbb{R}^d\backslash rB)\|g - f'_l\|^2_{L_2(\mu_{\mathbb{R}^d\backslash rB})} \\ &\le (1+\varepsilon)^2 \Bigl(e_i^2\bigl(B_H, L_2(\mu_{rB})\bigr) + r^{-\tau} e_i^2\bigl(B_H, L_2(\mu_{\mathbb{R}^d\backslash rB})\bigr)\Bigr)\,. \end{aligned}$$

Moreover, Carl's inequality (A.42) together with (A.27), (4.19), $\|k\|_\infty = 1$, and $a_m(S^*_{k,\nu}) = s_m(S^*_{k,\nu})$ for all $m \geq 1$ and all distributions ν on $\mathbb{R}^d$ yields a universal constant $K \geq 1$ independent of all other occurring terms such that

$$i e_i^2(B_H, L_2(\nu)) \leq \sum_{m=1}^{i} e_m^2(S^*_{k,\nu}) \leq K \|S^*_{k,\nu}\|^2_{\mathsf{HS}} \leq K\,. \tag{7.63}$$

Combining this estimate for $\nu := \mu_{\mathbb{R}^d \setminus rB}$ with $\sqrt{s+t} \leq \sqrt{s} + \sqrt{t}$ and the previous estimate, we then obtain

$$\|g - f_j \diamond f_l'\|_{L_2(\mu)} \leq (1+\varepsilon)\big(e_i(B_H, L_2(\mu_{rB})) + \sqrt{K} r^{-\tau/2} i^{-1/2}\big).$$

Since there are 2^{2i-2} functions $f_j \diamond f_l'$, the latter estimate together with (7.62) and $\varepsilon \to 0$ implies

$$e_{2i-1}(\mathrm{id} : H \to L_2(\mu)) \leq 2c\, a^\varsigma\, r^\varsigma i^{-\frac{1}{2p}} + 2\sqrt{K} r^{-\frac{\tau}{2}} i^{-\frac{1}{2}}\,,$$

where the factor 2 appears since in general we cannot guarantee $f_j \diamond f_l' \in H$ and hence (A.36) has to be applied. For $r := a^{-\frac{2\varsigma}{2\varsigma+\tau}} i^{\frac{1-p}{p(2\varsigma+\tau)}} \geq a^{-1}$, we thus arrive at

$$e_{2i-1}(\mathrm{id} : H \to L_2(\mu)) \leq 4c\sqrt{K}\, a^{\frac{\varsigma\tau}{2\varsigma+\tau}}\, i^{-\frac{2p\varsigma+\tau}{2p(2\varsigma+\tau)}}\,, \qquad i \geq 1,$$

and from this we easily obtain the assertion by the monotonicity of the entropy numbers. □

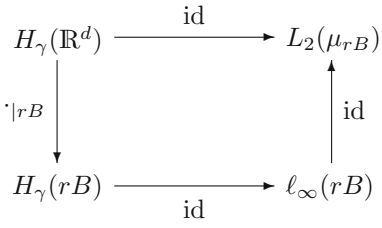
Let us now recall Example 4.32, where we saw that for all Gaussian RBF kernels the embedding $\mathrm{id} : H_\gamma(\mathbb{R}) \to C_b(\mathbb{R}^d)$ is *not* compact. Consequently, no entropy estimate of the form (7.49) is possible in this case. On the other hand, the following theorem together with Corollary 7.31 establishes an entropy assumption of the form (7.48), and hence the latter is indeed strictly weaker than (7.49).

Theorem 7.34 (Entropy numbers for Gaussian kernels). *Let μ be a distribution on $\mathbb{R}^d$ having tail exponent $\tau \in (0, \infty]$. Then, for all $\varepsilon > 0$ and $d/(d+\tau) < p < 1$, there exists a constant $c_{\varepsilon,p} \geq 1$ such that*

$$e_i\big(\mathrm{id} : H_\gamma(\mathbb{R}^d) \to L_2(\mu)\big) \leq c_{\varepsilon,p}\, \gamma^{-\frac{(1-p)(1+\varepsilon)d}{2p}}\, i^{-\frac{1}{2p}}$$

for all $i \geq 1$ and $\gamma \in (0, 1]$.

Proof. Let B be the closed unit Euclidean ball centered at the origin and $r > 0$ be a strictly positive real number. Since $\mathbb{R}^d \setminus rB$ has measure zero with respect to the distribution μ_{rB}, we obtain the commutative diagram



where $\cdot|_{rB}$ denotes the restriction operator. The latter is a metric surjection (note that Corollary 4.43 implies that it is even an isometric isomorphism), and, in addition, we obviously have $\|\operatorname{id} : \ell_\infty(rB) \to L_2(\mu_{rB})\| \le 1$. For an integer $m \ge 1$, Theorem 6.27 then yields a constant $\tilde{c}_{m,d} \ge 1$ only depending on m and d such that

$$e_i\big(\operatorname{id} : H_\gamma(\mathbb{R}^d) \to L_2(\mu_{rB})\big) \le e_i\big(\operatorname{id} : H_\gamma(rB) \to \ell_\infty(rB)\big) \le \tilde{c}_{m,d}\, r^m\, \gamma^{-m} i^{-\frac{m}{d}}$$

for all $0 < \gamma \le r$ and all $i \ge 1$. Let us now restrict our consideration to integers m with $m > d/2$. Applying Theorem 7.33 to the previous estimate, we then obtain

$$e_i\big(\operatorname{id} : H_\gamma(\mathbb{R}^d) \to L_2(\mu)\big) \;\le\; c_{m,d,\tau}\gamma^{-\frac{m\tau}{2m+\tau}}\, i^{-\frac{md+m\tau}{2md+d\tau}}\,, \tag{7.64}$$

where $c_{m,d,\tau} \ge 1$ is a constant only depending on m, d, and τ. Moreover, using (7.63) with $\nu := \mu$, we see that there exists a universal constant $K \ge 1$ such that

$$e_i\big(\operatorname{id} : H_\gamma(\mathbb{R}^d) \to L_2(\mu)\big) \;\le\; K\, i^{-1/2}\,, \qquad i \ge 1. \tag{7.65}$$

Interpolating (7.64) and (7.65) by Lemma A.1.3 with $t := 2$ and $r := \frac{2md+d\tau}{md+m\tau}$ shows that for all $s \in [r, 2]$ there exists a constant $c_{m,d,\tau,s} \ge 1$ such that

$$e_i\big(\operatorname{id} : H_\gamma(\mathbb{R}^d) \to L_2(\mathrm{D})\big) \;\le\; c_{m,d,\tau,s}\, \gamma^{-\frac{(2-s)md}{s(2m-d)}}\, i^{-\frac{1}{s}} \tag{7.66}$$

for all $i \ge 1$ and $m \ge 1$. We now fix a p with $\frac{d}{d+\tau} < p < 1$ and define $s := 2p$. For $\varepsilon > 0$, there then exists an integer m such that both

$$r = \frac{2md + d\tau}{md + m\tau} < s \qquad \text{and} \qquad \frac{md}{2m-d} \le \frac{(1+\varepsilon)d}{2}\,.$$

Consequently, we can apply (7.66) to this s and m. □

7.6 Further Reading and Advanced Topics

The first argument, presented in Section 7.1, for the *suboptimality* of the approach of Chapter 6 was discovered by Steinwart and Scovel (2005a, 2007), who also used the described iterative scheme to establish some learning rates for SVMs. The second argument, based on the variance bound for distributions

having zero Bayes risk, is well-known. For the case of binary classification, we refer to Section 12.7 of Devroye *et al.* (1996) for a detailed account including certain infinite sets of functions and some historical remarks.

In the machine learning literature, the idea of using a variance bound to obtain improved oracle inequalities goes, to the best of our knowledge, back to Lee *et al.* (1998). This idea was later refined by, e.g., Mendelson (2001a, 2001b), Bartlett *et al.* (2005), and Bartlett *et al.* (2006). In addition, similar ideas were also developed in the statistical literature by, e.g., Mammen and Tsybakov (1999) and Massart (2000b). For a brief historical survey, we refer to the introduction of Bartlett *et al.* (2005). The *improved oracle inequality for ERM* established in Theorem 7.2 is rooted in the ideas of the articles above. Finally, oracle inequalities that do *not* have a constant in front of the approximation error term have recently been established for ERM and some aggregation procedures by Lecué (2007a).

The first version of *Talagrand's inequality* was proved by Talagrand (1996), who showed that there exist universal constants K, c_1, and c_2 such that

$$\mathrm{P}\bigl(\{z \in Z : g(z) - \mathbb{E}_{\mathrm{P}} g \geq \varepsilon\}\bigr) \leq K \exp\left(-\frac{\varepsilon^2}{2(c_1 v + c_2 B\varepsilon)}\right), \tag{7.67}$$

where $v := \mathbb{E}_{\mathrm{P}} \sup_{f\in\mathcal{F}} \sum_{j=1}^n f(z_j)^2$. Unfortunately, however, his constants are rather large and, in addition, the variance condition expressed in v is more complicated then the one in Theorem A.9.1. The first improvement of this inequality was achieved by Ledoux (1996), who showed by developing the entropy functional technique that (7.67) holds for $K = 2$, $c_1 = 42$, and $c_2 = 8$ if v is replaced by $v + \frac{4}{21} B \mathbb{E}_{\mathrm{P}} g$. The next improvement was established by Massart (2000a) by proving (7.67) for $K = 1$, $c_1 = 8$, and $c_2 = 2.5$. In addition, he showed a version of Theorem 7.5 in which $\sqrt{2\tau n\sigma^2} + (2/3 + \gamma^{-1})\tau B$ is replaced by $\sqrt{8\tau n\sigma^2} + (5/2 + 32\gamma^{-1})\tau B$. The last improvements were then achieved by Rio (2002) and Bousquet (2002a). We followed Bousquet's approach, which gives optimal constants. Finally, Klein and Rio (2005) recently established a version of Talagrand's inequality for independent, not necessarily identically distributed random variables and almost optimal constants. More applications of the entropy technique are presented by Chafaï (2004), and other applications of Talagrand's inequality are discussed by Boucheron *et al.* (2003).

The *peeling* argument of Theorem 7.7 for weighted empirical processes is a standard tool to estimate complexity, measures of complicated function classes by complexity measures of related "easier" function classes. Moreover, *symmetrization* is a standard tool in empirical process theory that goes back to Kahane (1968) and Hoffmann-Jørgensen (1974, 1977). In the proof of Theorem A.8.1, we followed Section 2.3 of van der Vaart and Wellner (1996). The *contraction principle* stated in Theorem A.8.4 was taken from p. 112 of Ledoux and Talagrand (1991), and its Corollary A.8.5 was first shown by Talagrand (1994). Finally, *Dudley's chaining* (see Dudley, 1967, and the

historical remarks on p. 269 of van der Vaart and Wellner, 1996) for deriving the *maximal inequality* of Theorem 7.12 was taken from Section 2.2 of van der Vaart and Wellner (1996) with the additional improvements reported by Bousquet (2002b).

Empirical *Rademacher averages* were first used in learning theory by Koltchinskii (2001) and Koltchinskii and Panchenko (2000, 2002) as a penalization method for model selection. Some empirical findings for this approach were first reported by Lozano (2000). This penalization method was later refined, for example, by Lugosi and Wegkamp (2004). Furthermore, Rademacher averages of localized function classes were used by Bartlett *et al.* (2002), Mendelson (2002), and Bartlett *et al.* (2005) in a way similar to that of Theorem 7.20 and in a refined way by Bartlett *et al.* (2004). An overview of the ideas was described by Bousquet (2003b). Relations of Rademacher averages to other complexity measures are described by Mendelson (2002, 2003a), and structural properties of Rademacher averages were investigated by Bartlett and Mendelson (2002). Moreover, Mendelson (2003b) estimated Rademacher averages of balls in RKHSs by the eigenvalues of the associated integral operator. Finally, using covering numbers to bound Rademacher averages also goes back to Mendelson (2002). In Section 7.3, we essentially followed his approach, though we decided to use entropy numbers instead of covering numbers since *a)* this yields slightly smaller constants and *b)* entropy numbers have a conceptionally easier relation to eigenvalues and approximation numbers.

To the best of our knowledge, the *clipping idea* was first used by Bartlett (1998) for the analysis of neural networks. In the context of SVMs, it first appeared in a paper by Bousquet and Elisseeff (2002). The approach for the oracle inequalities presented in Section 7.4 was taken from Steinwart *et al.* (2007), which in turn was inspired by the work of Wu *et al.* (2007). An oracle inequality for SVMs using losses that *cannot* be clipped was established by Steinwart *et al.* (2006c). Their proof is based on a rather complicated refinement of a technique developed by Bartlett *et al.* (2006). However, it is relatively easy to modify the proof of the oracle inequality for CR-ERM presented in Theorem 7.20 to deal with regularized *unclipped* ERM. Such a modification would then essentially reproduce the oracle inequality of Steinwart *et al.* (2006c). Finally, an oracle inequality for unclipped SVMs using the hinge loss was proved by Blanchard *et al.* (2008). It is interesting to compare their conjecture regarding the optimal exponent in the regularization term with Exercise 7.6.

Unfortunately, little is known about the sharpness of the derived oracle inequalities and their resulting best learning rates. In fact, to the best of our knowledge, the only known results consider (essentially) the case where L is the least squares loss, $H = W^m(X)$ for some Euclidean ball X, P_X is the uniform distribution, and $m > d/2$. If the regression function $f^*_{L,\mathrm{P}}$ is bounded and contained in $W^k(X)$ for some $0 < k \leq m$, it is then easy to check by the remarks made in Section 5.6, the entropy number bound (A.48), and Corollary

7.31 that the best learning rate Theorem 7.23 provides is $n^{-\min\{\frac{k}{m}, \frac{2k}{2k+d}\}}$. For $m - d/2 \leq k \leq m$, it is known that this rate is optimal in a minmax sense. We refer to Györfi *et al.* (2002) and the references therein for such optimal rates.

Theorem 7.29, which compares the average eigenvalues of empirical integral operators with the eigenvalues of the corresponding infinite-sample integral operator, was first shown by Shawe-Taylor *et al.* (2002, 2005) in the special case of continuous kernels over compact metric spaces. Zwald *et al.* (2004) generalized this result to bounded measurable kernels with separable RKHSs. We essentially followed their ideas for the proof of Theorem 7.29, while to the best of our knowledge the rest of Section 7.5 has not been published.

Exercise 7.7 allows us to replace the entropy number assumption (7.48) by a bound on the eigenvalues of the integral operator. For the hinge loss and unclipped SVM decision functions, a conceptionally similar oracle inequality was shown by Blanchard *et al.* (2008). A key step in their proof is an estimate in the spirit of Mendelson (2003b), i.e., a bound of the Rademacher averages of balls of RKHSs by the eigenvalues of the associated integral operator. In the polynomial regime (7.68), their resulting learning rates are, however, always worse than the ones we obtained in the discussion after Theorem 7.23. This is, of course, partly a result of the fact that these authors consider unclipped decision functions. On the other hand, Steinwart and Scovel (2005b) pointed out that the results of Blanchard *et al.* (2008) are suboptimal for SVMs under the uniform entropy assumption (7.49), and by modifying the oracle inequality of Steinwart *et al.* (2006c) we conjecture that this remains true under the eigenvalue assumption (7.68). Finally, an oracle inequality for SVMs using the least squares loss that involves the eigenvalues of the associated integral operator was established by Caponnetto and De Vito (2005).

7.7 Summary

In this chapter, we developed advanced techniques for establishing oracle inequalities for ERM and SVMs. The main reason for this development was the observation made in Section 7.1 that the oracle inequalities of Chapter 6 provide suboptimal learning rates. Here we identified two sources for the suboptimality, namely a supremum bound that is suboptimal in terms of the regularization parameter and a possibly existing variance bound that can be exploited using Bernstein's inequality rather than Hoeffding's inequality.

In Section 7.2, we then established an improved oracle inequality for ERM over finite sets of functions that involves a variance bound. This oracle inequality proved to be useful for parameter selection purposes in Section 7.4. Moreover, its proof presented the core idea for establishing similar oracle inequalities for SVMs. Unfortunately, however, it turned out that this core idea could not be directly generalized to infinite sets of functions. This forced us to introduce some advanced tools from empirical process theory such as

Talagrand's inequality, peeling, symmetrization, Rademacher averages, and Dudley's chaining in Section 7.3. At the end of this section, we then illustrated how to orchestrate these tools to derive an improved oracle inequality for ERM over infinite sets of functions.

In Section 7.4, we used the tools from the previous section to establish an oracle inequality for general, *clipped regularized empirical risk minimizers*. We then derived two oracle inequalities for clipped SVMs, one involving bounds on entropy numbers and one not. An extensive comparison to the oracle inequalities of Chapter 6 showed that the new oracle inequalities always provide faster learning rates. Furthermore, we combined the improved oracle inequality for ERM over finite sets of functions with the new oracle inequalities for clipped SVMs to derive some results for simple data-dependent choices of the regularization parameter. Finally, we showed in Section 7.5 that the entropy assumptions of the new oracle inequalities can be substantially weaker than the ones from Chapter 6.

7.8 Exercises

7.1. Optimality of peeling ($\star$)
Let $T \neq \emptyset$ be a set, $g, h : T \to [0, \infty)$ be functions, and $r^* := \inf\{h(t) : t \in T\}$. Furthermore, let $\varphi : (r^*, \infty) \to [0, \infty)$ be a function such that

$$\sup_{\substack{t \in T \\ h(t) \leq r}} g(t) \;\geq\; \varphi(r)$$

for all $r > r^*$. Show the inequality

$$\sup_{t \in T} \frac{g(t)}{h(t) + r} \;\geq\; \frac{\varphi(r)}{2r}\,, \qquad\qquad r > r^*.$$

Generalize this result to expectations over suprema.

7.2. Simple properties of empirical Rademacher averages ($\star$)
Let $\mathcal{F}, \mathcal{G} \subset \mathcal{L}_0(Z)$ be non-empty subsets, $\alpha \in \mathbb{R}$ be a real number, n be an integer, and $D := (z_1, \ldots, z_n) \in Z^n$ be a finite sequence. Show that the following relations hold:

$$\begin{aligned}
\mathrm{Rad}_D(\mathcal{F} \cup \mathcal{G}, n) &\leq \mathrm{Rad}_D(\mathcal{F}, n) + \mathrm{Rad}_D(\mathcal{G}, n)\,,\\
\mathrm{Rad}_D(\mathcal{F} \cup \{0\}, n) &= \mathrm{Rad}_D(\mathcal{F}, n)\,,\\
\mathrm{Rad}_D(\alpha\mathcal{F}, n) &= |\alpha| \mathrm{Rad}_D(\mathcal{F}, n)\,,\\
\mathrm{Rad}_D(\mathcal{F} + \mathcal{G}, n) &\leq \mathrm{Rad}_D(\mathcal{F}, n) + \mathrm{Rad}_D(\mathcal{G}, n)\,,\\
\mathrm{Rad}_D(\mathrm{co}\,\mathcal{F}, n) &= \mathrm{Rad}_D(\mathcal{F}, n)\,,
\end{aligned}$$

where in the last equation $\mathrm{co}\,\mathcal{F}$ denotes the convex hull of $\mathcal{F}$.

7.3. Another oracle inequality ($\star\star\star$)
Let $L : X \times Y \times \mathbb{R} \to [0, \infty)$ be a Lipschitz continuous loss with $|L|_1 \leq 1$, $\mathcal{F} \subset \mathcal{L}_\infty(X)$ be a separable set satisfying $\|f\|_\infty \leq M$ for a suitable constant $M > 0$ and all $f \in \mathcal{F}$, and P be a distribution on $X \times Y$ that has a Bayes decision function $f^*_{L,\mathrm{P}}$ with $\mathcal{R}_{L,\mathrm{P}}(f^*_{L,\mathrm{P}}) < \infty$. Assume that there exist constants $B > 0$, $\vartheta \in [0, 1]$, and $V \geq B^{2-\vartheta}$ such that for all $f \in \mathcal{F}$ we have

$$\begin{aligned} \|L \circ f - L \circ f^*_{L,\mathrm{P}}\|_\infty &\leq B\,, \\ \mathbb{E}_\mathrm{P}\bigl(L \circ f - L \circ f^*_{L,\mathrm{P}}\bigr)^2 &\leq V \cdot \bigl(\mathbb{E}_\mathrm{P}(L \circ f - L \circ f^*_{L,\mathrm{P}})\bigr)^\vartheta\,. \end{aligned}$$

For $f \in \mathcal{F}$, define $h_f := L \circ f - L \circ f^*_{L,\mathrm{P}}$. For a measurable ERM with respect to L and $\mathcal{F}$, show the following assertions:

i) $\|h_f - h_{f'}\|_\infty \leq \|f - f'\|_\infty$ for all $f, f' \in \mathcal{F}$.
ii) Let $\mathcal{C}$ be an ε-net of $\mathcal{F}$ with respect to $\|\cdot\|_\infty$. Then for all $D \in (X \times Y)^n$ there exists an $f \in \mathcal{C}$ such that $\mathbb{E}_\mathrm{P} h_{f_D} - \mathbb{E}_\mathrm{D} h_{f_D} \leq \mathbb{E}_\mathrm{P} h_f - \mathbb{E}_\mathrm{D} h_f + 2\varepsilon$.
iii) For all $\varepsilon > 0$, $r > 0$, and $\tau > 0$, we have

$$\mathbb{E}_\mathrm{P} h_{f_D} - \mathbb{E}_\mathrm{D} h_{f_D} < 2\varepsilon + \bigl(\mathbb{E}_\mathrm{P} h_{f_D} + \varepsilon\bigr)\left(\sqrt{\frac{2V\tau}{nr^{2-\vartheta}}} + \frac{4B\tau}{3nr}\right) + \sqrt{\frac{2V\tau r^\vartheta}{n}} + \frac{4B\tau}{3n}$$

with probability P^n not less than $1 - \mathcal{N}(\mathcal{F}, \|\cdot\|_\infty, \varepsilon)e^{-\tau}$.
iv) Given an $\varepsilon > 0$ and $\tau > 0$, we have

$$\begin{aligned} \mathcal{R}_{L,\mathrm{P}}(f_D) - \mathcal{R}^*_{L,\mathrm{P}} \leq\, & 6\bigl(\mathcal{R}^*_{L,\mathrm{P},\mathcal{F}} - \mathcal{R}^*_{L,\mathrm{P}}\bigr) + 8\varepsilon \\ & + 4\left(\frac{8V\bigl(\tau + 1 + \ln \mathcal{N}(\mathcal{F}, \|\cdot\|_\infty, \varepsilon)\bigr)}{n}\right)^{\frac{1}{2-\vartheta}} \end{aligned}$$

with probability P^n not less than $1 - e^{-\tau}$.
v) Let $a > 0$ and $p > 0$ be constants such that $\ln \mathcal{N}(\mathcal{F}, \|\cdot\|_\infty, \varepsilon) \leq a\varepsilon^{-2p}$ for all $\varepsilon > 0$. Then, for all $\tau \geq 1$, we have

$$\mathcal{R}_{L,\mathrm{P}}(f_D) - \mathcal{R}^*_{L,\mathrm{P}} \leq 6\bigl(\mathcal{R}^*_{L,\mathrm{P},\mathcal{F}} - \mathcal{R}^*_{L,\mathrm{P}}\bigr) + 4\left(\frac{16V\tau}{n}\right)^{\frac{1}{2-\vartheta}} + c_{p,\vartheta}\left(\frac{8Va}{n}\right)^{\frac{1}{2+2p-\vartheta}}$$

with probability P^n not less than $1 - e^{-\tau}$, where $c_{p,\vartheta} \in (0, 12]$ is a constant with $\lim_{p \to 0^+} c_{p,\vartheta} = 4$.
vi) Compare this oracle inequality with (7.28) and (7.30).

Hint: To prove *iii)*, fix an ε-net $\mathcal{C}$ of $\mathcal{F}$ with $|\mathcal{C}| = \mathcal{N}(\mathcal{F}, \|\cdot\|_\infty, \varepsilon)$. Then apply (7.9) to $\mathcal{C}$. Finally, follow the argument after (7.9) and apply *ii)*. To prove *iv)*, repeat the last steps of the proof of Theorem 7.2 using *iii)* for $r := (\frac{16V\tau}{n})^{1/(2-\vartheta)}$.

7.4. Oracle inequality for clipped approximate SVMs ($\star\star\star$)
Under the assumptions of Theorem 7.23, show that any clipped ϵ-approximate SVM satisfies

$$\mathrm{P}^n\Bigl(D : \lambda\|f_{\mathrm{D},\lambda}\|^2+\mathcal{R}_{L,\mathrm{P}}(\widehat{f}_{\mathrm{D},\lambda})-\mathcal{R}^*_{L,\mathrm{P}} > 6\bigl(\mathcal{R}^{reg}_{L,\mathrm{P},\lambda}(f_0)-\mathcal{R}^*_{L,\mathrm{P}}\bigr)+3r+3\epsilon\Bigr) \le 3e^{-\tau}.$$

7.5. Comparison of oracle inequalities ($\star\star\star$)
Apply Theorem 7.20 to the assumptions of Exercise 7.3, and show that Theorem 7.20 produces an oracle inequality that is sharper in n.

7.6. Different exponents in the regularization term of SVMs ($\star\star\star\star$)
Assume that L, H, P, and f_0 as well as M, B, V, ϑ, a, p, and B_0 are as in Theorem 7.23. Moreover, for $q \in [0,\infty)$ and $\lambda > 0$, consider the learning method that assigns to every $D \in (X \times Y)^n$ a function $f^{(q)}_{\mathrm{D},\lambda} \in H$ that solves

$$\min_{f\in H} \lambda\|f\|_H^q + \mathcal{R}_{L,\mathrm{D}}(f)\,.$$

Show that for all fixed $\tau > 0$, $n \ge 1$, and $\lambda > 0$, this learning method satisfies

$$\begin{aligned}\lambda\|f^{(q)}_{\mathrm{D},\lambda}\|_H^q+\mathcal{R}_{L,\mathrm{P}}(\widehat{f}^{(q)}_{\mathrm{D},\lambda})-\mathcal{R}^*_{L,\mathrm{P}} \le{}& 9\bigl(\lambda\|f_0\|_H^q + \mathcal{R}_{L,\mathrm{P}}(f_0) - \mathcal{R}^*_{L,\mathrm{P}}\bigr) + \frac{15B_0\tau}{n}\\ &+K\Bigl(\frac{a^{2pq}}{\lambda^{2p}n^q}\Bigr)^{\frac{1}{2q-2p-\vartheta q+\vartheta pq}}+3\Bigl(\frac{72V\tau}{n}\Bigr)^{\frac{1}{2-\vartheta}}\end{aligned}$$

with probability P^n not less than $1-3e^{-\tau}$, where $K \ge 1$ is a constant only depending on p, M, B, ϑ, and V.

Furthermore, assume that there exist constants $c > 0$ and $\beta \in (0,1]$ such that $A_2(\lambda) \le c\lambda^\beta$ for all $\lambda \ge 0$. In addition, suppose that L is Lipschitz continuous. Using Exercise 5.11 and Lemma A.1.7, show that (7.53) are the best learning rates this oracle inequality can provide. Can we make a conclusion regarding the "optimal" exponent q?

7.7. Eigenvalues vs. average entropy numbers ($\star\star\star$)
Let k be a measurable kernel on X with separable RKHS H and μ be a probability measure on X such that $\|k\|_{L_2(\mu)} < \infty$. Assume that there exist constants $0 < p < 1$ and $a \ge 1$ such that the extended sequence of eigenvalues of the integral operator $T_{k,\mu}$ satisfies

$$\lambda_i(T_{k,\mu}) \le a\, i^{-\frac{1}{p}}\,, \qquad i \ge 1. \tag{7.68}$$

Show that there exists a constant $c_p > 0$ only depending on p such that

$$\mathbb{E}_{D\sim\mu^n} e_i(B_H, L_2(\mathrm{D})) \le c_p\, a\, \bigl(\min\{i,n\}\bigr)^{\frac{1}{2p}}\, i^{-\frac{1}{p}}\,, \qquad i,n \ge 1.$$