5

Infinite-Sample Versions of Support Vector Machines

Overview. *In this chapter, we show that interesting structural properties of SVMs can be discovered by considering the SVM formulation for "infinite" training sets. In particular, we will present a representation for the corresponding SVM solutions and discuss their dependence on the underlying probability measure. Moreover, we will investigate the behavior of SVMs for vanishing regularization parameter.*

Prerequisites. *Chapters 2 and 4 form the fundament of this chapter. In addition, we need the subdifferential calculus for convex functions, which is provided in Section A.6.2.*

Usage. *Section 5.1 is needed for computational aspects discussed in Chapter 11, while Section 5.2 is mainly used in Section 5.3. The latter section is required for the statistical analysis of SVMs conducted in Sections 6.4, 6.5, and 9.2. Finally, Sections 5.4 and 5.5 are important for the generalization performance of SVMs analyzed in Sections 6.4 and 6.5 and Chapters 8 and 9.*

We saw in the introduction that support vector machines obtain their decision functions by finding a minimizer $f_{\mathrm{D},\lambda}$ of the regularized empirical risk

$$\mathcal{R}^{reg}_{L,\mathrm{D},\lambda}(f) := \lambda\|f\|^2_H + \mathcal{R}_{L,\mathrm{D}}(f)\,, \qquad\qquad f \in H\,. \tag{5.1}$$

The first questions that then arise are whether such minimizers exist and, if so, whether they are unique. Moreover, the data set D defining the empirical measure D is usually an i.i.d. sample drawn from a distribution P. The law of large numbers then shows that $\mathcal{R}_{L,\mathrm{D}}(f)$ is close to $\mathcal{R}_{L,\mathrm{P}}(f)$, and hence one may think of $\mathcal{R}^{reg}_{L,\mathrm{D},\lambda}(f)$ as an estimate of the *infinite-sample* regularized risk

$$\mathcal{R}^{reg}_{L,\mathrm{P},\lambda}(f) := \lambda\|f\|^2_H + \mathcal{R}_{L,\mathrm{P}}(f)\,. \tag{5.2}$$

In this chapter, we will thus investigate such regularized risks for *arbitrary* distributions, so that we simultaneously obtain results for (5.1) and (5.2). In particular, we will consider the following questions:

- When does (5.2) have exactly one minimizer $f_{\mathrm{P},\lambda} \in H$?
- Is there a way to represent $f_{\mathrm{P},\lambda}$ in a suitable form?
- How does $f_{\mathrm{P},\lambda}$ change if P or λ changes?
- How close is $\mathcal{R}_{L,\mathrm{P}}(f_{\mathrm{P},\lambda})$ to the Bayes risk $\mathcal{R}^*_{L,\mathrm{P}}$?

Namely, we show in Section 5.1 that under rather general conditions there exists a unique minimizer $f_{\mathrm{P},\lambda}$, and in the following section we derive a representation for this minimizer. This representation is then used in Section 5.3 to investigate the dependence of $f_{\mathrm{P},\lambda}$ on P. In Sections 5.4 and 5.5, we finally consider the question of whether and how $\mathcal{R}_{L,\mathrm{P}}(f_{\mathrm{P},\lambda})$ converges to the Bayes risk for $\lambda \to 0$.

5.1 Existence and Uniqueness of SVM Solutions

In this section, we first investigate under which conditions the general regularized risk (5.2) possesses a unique minimizer. Furthermore, we establish a representation of the finite sample solutions $f_{\mathrm{D},\lambda}$ of (5.1).

Let us begin by introducing some notions. To this end, let $L : X \times Y \times \mathbb{R} \to [0, \infty)$ be a loss, H be the RKHS of a measurable kernel on X, and P be a distribution on $X \times Y$. For $\lambda > 0$, we then call an $f_{\mathrm{P},\lambda,H} \in H$ that satisfies

$$\lambda \|f_{\mathrm{P},\lambda,H}\|_H^2 + \mathcal{R}_{L,\mathrm{P}}(f_{\mathrm{P},\lambda,H}) = \inf_{f \in H} \lambda \|f\|_H^2 + \mathcal{R}_{L,\mathrm{P}}(f) \tag{5.3}$$

a **general SVM solution** or a **general SVM decision function**. Moreover, in order to avoid notational overload, we usually use the shorthand $f_{\mathrm{P},\lambda} := f_{\mathrm{P},\lambda,H}$ if no confusion can arise. Now note that for such a function $f_{\mathrm{P},\lambda}$ we have

$$\lambda \|f_{\mathrm{P},\lambda}\|_H^2 \leq \lambda \|f_{\mathrm{P},\lambda}\|_H^2 + \mathcal{R}_{L,\mathrm{P}}(f_{\mathrm{P},\lambda}) = \inf_{f \in H} \lambda \|f\|_H^2 + \mathcal{R}_{L,\mathrm{P}}(f) \leq \mathcal{R}_{L,\mathrm{P}}(0)\,,$$

or in other words

$$\|f_{\mathrm{P},\lambda}\|_H \leq \sqrt{\frac{\mathcal{R}_{L,\mathrm{P}}(0)}{\lambda}}\,. \tag{5.4}$$

Let us now investigate under which assumptions there exists exactly one $f_{\mathrm{P},\lambda}$. We begin with the following result showing uniqueness for convex losses.

Lemma 5.1 (Uniqueness of SVM solutions). *Let $L : X \times Y \times \mathbb{R} \to [0, \infty)$ be a convex loss, H be the RKHS of a measurable kernel over X, and* P *be a distribution on $X \times Y$ with $\mathcal{R}_{L,\mathrm{P}}(f) < \infty$ for some $f \in H$. Then for all $\lambda > 0$ there exists at most one general SVM solution $f_{\mathrm{P},\lambda}$.*

Proof. Let us assume that the map $f \mapsto \lambda \|f\|_H^2 + \mathcal{R}_{L,\mathrm{P}}(f)$ has two minimizers $f_1, f_2 \in H$ with $f_1 \neq f_2$. By the last statement in Lemma A.5.9, we then find $\|\frac{1}{2}(f_1 + f_2)\|_H^2 < \frac{1}{2}\|f_1\|_H^2 + \frac{1}{2}\|f_2\|_H^2$. The convexity of $f \mapsto \mathcal{R}_{L,\mathrm{P}}(f)$ together with $\lambda \|f_1\|_H^2 + \mathcal{R}_{L,\mathrm{P}}(f_1) = \lambda \|f_2\|_H^2 + \mathcal{R}_{L,\mathrm{P}}(f_2)$ then shows for $f^* := \frac{1}{2}(f_1 + f_2)$ that

$$\lambda \|f^*\|_H^2 + \mathcal{R}_{L,\mathrm{P}}(f^*) < \lambda \|f_1\|_H^2 + \mathcal{R}_{L,\mathrm{P}}(f_1)\,,$$

i.e., f_1 is not a minimizer of $f \mapsto \lambda \|f\|_H^2 + \mathcal{R}_{L,\mathrm{P}}(f)$. Consequently, the assumption that there are two minimizers is false. □

Our next result shows that for convex, integrable Nemitski loss functions (see Definition 2.16) there always exists a general SVM solution.

Theorem 5.2 (Existence of SVM solutions). *Let $L : X \times Y \times \mathbb{R} \to [0, \infty)$ be a convex loss and* P *be a distribution on $X \times Y$ such that L is a* P*-integrable Nemitski loss. Furthermore, let H be the RKHS of a bounded measurable kernel over X. Then, for all $\lambda > 0$, there exists a general SVM solution $f_{\mathrm{P},\lambda}$.*

Proof. Since the kernel k of H is measurable, H consists of measurable functions by Lemma 4.24. Moreover, k is bounded, and thus Lemma 4.23 shows that $\mathrm{id} : H \to L_\infty(\mathrm{P}_X)$ is continuous. In addition, we have $L(x, y, t) < \infty$ for all $(x, y, t) \in X \times Y \times \mathbb{R}$, and hence L is a continuous loss by the convexity of L and Lemma A.6.2. Therefore, Lemma 2.17 shows that $\mathcal{R}_{L,\mathrm{P}} : L_\infty(\mathrm{P}_X) \to \mathbb{R}$ is continuous, and hence $\mathcal{R}_{L,\mathrm{P}} : H \to \mathbb{R}$ is continuous. In addition, Lemma 2.13 provides the convexity of this map. Furthermore, $f \mapsto \lambda \|f\|_H^2$ is also convex and continuous, and hence so is $f \mapsto \lambda \|f\|_H^2 + \mathcal{R}_{L,\mathrm{P}}(f)$. Now consider the set

$$A := \left\{ f \in H : \lambda \|f\|_H^2 + \mathcal{R}_{L,\mathrm{P}}(f) \le M \right\},$$

where $M := \mathcal{R}_{L,\mathrm{P}}(0)$. Then we obviously have $0 \in A$. In addition, $f \in A$ implies $\lambda \|f\|_H^2 \le M$, and hence $A \subset (M/\lambda)^{1/2} B_H$, where B_H is the closed unit ball of H. In other words, A is a non-empty and bounded subset and thus Theorem A.6.9 gives the existence of a minimizer $f_{\mathrm{P},\lambda}$. □

It is interesting to note that in Theorem 5.2 the convexity of L is *not necessary* for the existence of a general SVM solution (see Exercise 5.7).

We have presented some important classes of Nemitski losses in Chapter 2. The following corollaries specify Theorem 5.2 for these types of losses.

Corollary 5.3. *Let $L : X \times Y \times \mathbb{R} \to [0, \infty)$ be a convex, locally Lipschitz continuous loss,* P *be a distribution on $X \times Y$ with $\mathcal{R}_{L,\mathrm{P}}(0) < \infty$, and H be the RKHS of a bounded measurable kernel over X. Then, for all $\lambda > 0$, there exists a unique general SVM solution $f_{\mathrm{P},\lambda} \in H$. Furthermore, if L is actually a convex margin-based loss represented by $\varphi : \mathbb{R} \to [0, \infty)$, we have*

$$\|f_{\mathrm{P},\lambda}\|_H \;\le\; \left(\frac{\varphi(0)}{\lambda} \right)^{1/2}.$$

Proof. The first assertion follows from Lemma 5.1, Theorem 5.2, and the discussion around (2.11). Moreover, convex margin-based losses are locally Lipschitz continuous by Lemma 2.25, and (5.4) together with $\mathcal{R}_{L,\mathrm{P}}(0) = \varphi(0)$ shows the inequality for these losses. □

Corollary 5.4. *Let $L : \mathbb{R} \times \mathbb{R} \to [0, \infty)$ be a convex, distance-based loss of upper growth type $p \ge 1$,* P *be a distribution on $X \times Y$ with $|\mathrm{P}|_p < \infty$, and H be the RKHS of a bounded measurable kernel over X. Then, for all $\lambda > 0$, there exists a unique general SVM solution $f_{\mathrm{P},\lambda} \in H$. Moreover, there exists a constant $c_{L,p} > 0$ only depending on L and p such that*

$$\|f_{\mathrm{P},\lambda}\|_H \;\le\; c_{L,p} \left(\frac{|\mathrm{P}|_p^p + 1}{\lambda} \right)^{1/2}.$$

Proof. Combine Lemma 2.38, Lemma 5.1, Theorem 5.2, and (5.4). □

If H is the RKHS of a bounded measurable kernel on X and L is a convex, distance-based loss of growth type $p \geq 1$, then it is easy to see by Lemma 2.38 that, for distributions P on $X \times \mathbb{R}$ with $|\mathrm{P}|_p = \infty$, we have $\mathcal{R}_{L,\mathrm{P}}(f) = \infty$ for *all* $f \in H$. Consequently, the distributions P with $|\mathrm{P}|_p < \infty$ are in general the only distributions admitting a *non-trivial* SVM solution $f_{\mathrm{P},\lambda}$.

Let us now discuss **empirical SVM solutions** in some more detail. To this end, we denote, as usual, the empirical measure of a sequence of observations $D := ((x_1, y_1), \dots, (x_n, y_n))$ by D, i.e., $\mathrm{D} := \frac{1}{n}\sum_{i=1}^n \delta_{(x_i,y_i)}$. The next result shows that $f_{\mathrm{D},\lambda}$ exists under somewhat minimal assumptions on L. Furthermore, it provides a simple representation of $f_{\mathrm{D},\lambda}$ in terms of the kernel.

Theorem 5.5 (Representer theorem). *Let $L : X \times Y \times \mathbb{R} \to [0, \infty)$ be a convex loss and $D := ((x_1, y_1), \dots, (x_n, y_n)) \in (X \times Y)^n$. Furthermore, let H be an RKHS over X. Then, for all $\lambda > 0$, there exists a unique empirical SVM solution, i.e., a unique $f_{\mathrm{D},\lambda} \in H$ satisfying*

$$\lambda \|f_{\mathrm{D},\lambda}\|_H^2 + \mathcal{R}_{L,\mathrm{D}}(f_{\mathrm{D},\lambda}) = \inf_{f \in H} \lambda \|f\|_H^2 + \mathcal{R}_{L,\mathrm{D}}(f) . \tag{5.5}$$

In addition, there exist $\alpha_1, \dots, \alpha_n \in \mathbb{R}$ such that

$$f_{\mathrm{D},\lambda}(x) = \sum_{i=1}^n \alpha_i k(x, x_i) , \qquad x \in X. \tag{5.6}$$

Proof. Uniqueness. It follows by repeating the proof of Lemma 5.1.

Existence. Since convergence in H implies pointwise convergence, we obtain the continuity of $\mathcal{R}_{L,\mathrm{D}} : H \to [0, \infty)$ by the continuity of L. Now the existence can be shown as in the proof of Theorem 5.2.

Representation (5.6). Let us write $X' := \{x_i : i = 1, \dots, n\}$ and

$$H_{|X'} := \operatorname{span}\bigl\{k(\,\cdot\,, x_i) : i = 1, \dots, n\bigr\} .$$

Then $H_{|X'}$ is the RKHS of $k_{|X' \times X'}$ by (4.12), and consequently we already know that there exists an empirical SVM solution $f_{\mathrm{D},\lambda,H_{|X'}} \in H_{|X'}$. Now let $H_{|X'}^{\perp}$ be the orthogonal complement of $H_{|X'}$ in H. For $f \in H_{|X'}^{\perp}$, we then have

$$f(x_i) = \langle f, k(\,\cdot\,, x_i)\rangle = 0 , \qquad i = 1, \dots, n .$$

If $P_{X'} : H \to H$ denotes the orthogonal projection onto $H_{|X'}$ we thus find

$$\mathcal{R}_{L,\mathrm{D}}(P_{X'}f) = \mathcal{R}_{L,\mathrm{D}}(f)$$

and $\|P_{X'}f\|_H \leq \|f\|_H$ for all $f \in H$. Since this yields

$$\begin{aligned}
\inf_{f \in H} \lambda \|f\|_H^2 + \mathcal{R}_{L,\mathrm{D}}(f) &\leq \inf_{f \in H_{|X'}} \lambda \|f\|_H^2 + \mathcal{R}_{L,\mathrm{D}}(f) \\
&= \inf_{f \in H} \lambda \|P_{X'}f\|_H^2 + \mathcal{R}_{L,\mathrm{D}}(P_{X'}f) \\
&\leq \inf_{f \in H} \lambda \|f\|_H^2 + \mathcal{R}_{L,\mathrm{D}}(f) ,
\end{aligned}$$

we actually have equality in the above chain of inequalities. Therefore, $f_{\mathrm{D},\lambda,H_{|X'}}$ minimizes $\lambda \mapsto \lambda\|f\|_H^2 + \mathcal{R}_{L,\mathrm{D}}(f)$ in H, and the uniqueness of $f_{\mathrm{D},\lambda,H}$ then shows $f_{\mathrm{D},\lambda,H_{|X'}} = f_{\mathrm{D},\lambda,H}$. In other words, we have $f_{\mathrm{D},\lambda} := f_{\mathrm{D},\lambda,H} \in H_{|X'}$, and hence (5.6) follows from the definition of $H_{|X'}$. □

Our last theorem in this section shows that in most situations the general SVM decision functions are non-trivial.

Theorem 5.6 (Non-trivial SVM solutions). *Let $L : X \times Y \times \mathbb{R} \to [0,\infty)$ be a convex loss and P be a distribution on $X \times Y$ such that L is a P-integrable Nemitski loss. Furthermore, let H be the RKHS of a bounded measurable kernel over X such that $\mathcal{R}^*_{L,\mathrm{P},H} < \mathcal{R}_{L,\mathrm{P}}(0)$. Then, for all $\lambda > 0$, we have $f_{\mathrm{P},\lambda} \neq 0$.*

Proof. By our assumptions, there exists an $f^* \in H$ with $\mathcal{R}_{L,\mathrm{P}}(f^*) < \mathcal{R}_{L,\mathrm{P}}(0)$. For $\alpha \in [0,1]$, we then have

$$\lambda\|\alpha f^*\|_H^2 + \mathcal{R}_{L,\mathrm{P}}(\alpha f^*) \le \lambda\alpha^2\|f^*\|_H^2 + \alpha\mathcal{R}_{L,\mathrm{P}}(f^*) + (1-\alpha)\mathcal{R}_{L,\mathrm{P}}(0) =: h(\alpha)$$

by the convexity of $\mathcal{R}_{L,\mathrm{P}}$. Now, $\mathcal{R}_{L,\mathrm{P}}(f^*) < \mathcal{R}_{L,\mathrm{P}}(0)$ together with the quadratic form of $\alpha \mapsto h(\alpha)$ implies that $h : [0,1] \to [0,\infty)$ is minimized at some $\alpha^* \in (0,1]$. Consequently, we have

$$\lambda\|\alpha^* f^*\|_H^2 + \mathcal{R}_{L,\mathrm{P}}(\alpha^* f^*) \le h(\alpha^*) < h(0) = \lambda\|0\|_H^2 + \mathcal{R}_{L,\mathrm{P}}(0)\,. \qquad \square$$

5.2 A General Representer Theorem

We have seen in Theorem 5.5 that *empirical* SVM solutions can be represented by linear combinations of the canonical feature map. However, this result does not provide information about the *values* of the coefficients $\alpha_1, \ldots, \alpha_n$, and, in addition, it also remains unclear whether a somewhat similar result can hold for *general* SVM solutions.

Let us begin with a simplified consideration. To this end, let X be a measurable space, P be a distribution on $X \times Y$, and $L : X \times Y \times \mathbb{R} \to [0,\infty)$ be a convex, differentiable, and P-integrable Nemitski loss. Furthermore, assume that $|L'| : X \times Y \times \mathbb{R} \to [0,\infty)$ is also a P-integrable Nemitski loss. Then Lemma 2.21 shows that $\mathcal{R}_{L,\mathrm{P}} : L_\infty(\mathrm{P}_X) \to [0,\infty)$ is Fréchet differentiable and that its derivative at $f \in L_\infty(\mathrm{P}_X)$ is the bounded linear operator $\mathcal{R}'_{L,\mathrm{P}}(f) : L_\infty(\mathrm{P}_X) \to \mathbb{R}$ given by

$$\mathcal{R}'_{L,\mathrm{P}}(f)g = \int_{X\times Y} g(x)L'\big(x,y,f(x)\big)\,d\mathrm{P}(x,y)\,, \qquad g \in L_\infty(\mathrm{P}_X).$$

Now let H be a separable RKHS with bounded and measurable kernel k and $\Phi : X \to H$ be the corresponding canonical feature map. Lemma 4.25 then shows that $\Phi : X \to H$ is measurable, and from Lemmas 4.23 and 4.24 we can infer that $\mathrm{id} : H \to L_\infty(\mathrm{P}_X)$ is well-defined and continuous. For $f_0 \in H$, the chain rule (see Lemma A.5.15) thus yields

$$\big(\mathcal{R}_{L,\mathrm{P}} \circ \mathrm{id}\big)'(f_0) = \mathcal{R}'_{L,\mathrm{P}}(\mathrm{id}\, f_0) \circ \mathrm{id}'(f_0) = \mathcal{R}'_{L,\mathrm{P}}(f_0) \circ \mathrm{id}\,,$$

and hence we find for $f \in H$ that

$$\begin{aligned}\big(\mathcal{R}_{L,\mathrm{P}} \circ \mathrm{id}\big)'(f_0)f = \big(\mathcal{R}'_{L,\mathrm{P}}(f_0) \circ \mathrm{id}\big)f &= \int_{X\times Y} f(x)L'\big(x,y,f_0(x)\big)\, d\mathrm{P}(x,y) \\ &= \mathbb{E}_{(x,y)\sim\mathrm{P}} L'\big(x,y,f_0(x)\big)\langle f, \Phi(x)\rangle \\ &= \Big\langle f, \mathbb{E}_{(x,y)\sim\mathrm{P}} L'\big(x,y,f_0(x)\big)\Phi(x)\Big\rangle\,.\end{aligned}$$

Note that the last expectation is H-valued and hence a *Bochner integral* in the sense of Section A.5.4. Using the Fréchet-Riesz isomorphism $\iota : H \to H'$ described in Theorem A.5.12, we thus see that

$$\big(\mathcal{R}_{L,\mathrm{P}} \circ \mathrm{id}\big)'(f_0) = \iota\mathbb{E}_{(x,y)\sim\mathrm{P}} L'\big(x,y,f_0(x)\big)\Phi(x)\,. \tag{5.7}$$

Moreover, a straightforward calculation shows that the function $G : H \to \mathbb{R}$ defined by $Gf := \|f\|_H^2$, $f \in H$, is Fréchet differentiable and its derivative at f_0 is $G'(f_0) = 2\iota f_0$. Now recall that $\mathcal{R}^{reg}_{L,\mathrm{P},\lambda}(\,\cdot\,) = \lambda G + \mathcal{R}_{L,\mathrm{P}} \circ \mathrm{id}$, and hence $f_{\mathrm{P},\lambda}$ minimizes the function $\lambda G + \mathcal{R}_{L,\mathrm{P}} \circ \mathrm{id} : H \to \mathbb{R}$. Consequently, we obtain

$$0 = \big(\lambda G + \mathcal{R}_{L,\mathrm{P}} \circ \mathrm{id}\big)'(f_{\mathrm{P},\lambda}) = \iota\Big(2\lambda f_{\mathrm{P},\lambda} + \mathbb{E}_{(x,y)\sim\mathrm{P}} L'\big(x,y,f_{\mathrm{P},\lambda}(x)\big)\Phi(x)\Big)\,,$$

and hence we have $2\lambda f_{\mathrm{P},\lambda} = -\mathbb{E}_{(x,y)\sim\mathrm{P}} L'(x,y,f_{\mathrm{P},\lambda}(x))\Phi(x)$ by the injectivity of ι. The reproducing property then yields

$$f_{\mathrm{P},\lambda}(x) = -\int_{X\times Y} \frac{L'(x',y,f_{\mathrm{P},\lambda}(x'))}{2\lambda} k(x,x')\, d\mathrm{P}(x',y)\,, \qquad x \in X. \tag{5.8}$$

This equation answers both questions we posed in the introduction of this section. Indeed, for a data set $D = ((x_1,y_1),\dots,(x_n,y_n))$ with corresponding empirical distribution D, equation (5.8) becomes

$$f_{\mathrm{D},\lambda}(x) = -\frac{1}{2\lambda n}\sum_{i=1}^{n} L'\big(x_i,y_i,f_{\mathrm{D},\lambda}(x_i)\big)k(x,x_i)\,, \qquad x \in X,$$

i.e., possible coefficients in (5.6) are

$$\alpha_i := -\frac{L'(x_i,y_i,f_{\mathrm{D},\lambda}(x_i))}{2\lambda n}\,, \qquad i = 1,\dots,n.$$

Moreover, note that by writing $h(x,y) := L'(x,y,f_{\mathrm{D},\lambda}(x))$, $(x,y) \in X \times Y$, the representation in (5.6) becomes

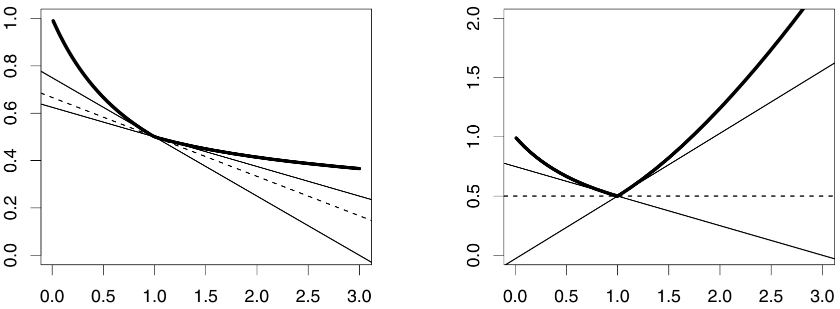

Fig. 5.1. Convex functions (bold lines) and some of their subdifferentials. Left: The function f is not differentiable at $x := 1$, so that the subdifferential $\partial f(x)$ contains the slopes of all tangents of f at x. In particular, it contains the left and the right derivatives of f at x (solid lines). Moreover, all slopes in between (dashed line) are contained. A formal statement of this illustration can be found in Lemma A.6.15. Right: The subdifferential at a minimum contains 0, i.e., the flat slope (dashed line).

$$f_{\mathrm{D},\lambda} = -\frac{1}{2\lambda n}\sum_{i=1}^{n} h(x_i, y_i) k(\,\cdot\,, x_i) = -\frac{1}{2\lambda}\mathbb{E}_{\mathrm{D}} h\Phi\,.$$

On the other hand, (5.8) can be rewritten as

$$f_{\mathrm{P},\lambda} = -\frac{1}{2\lambda}\int_{X\times Y} h(x,y)k(\,\cdot\,,x)\,d\mathrm{P}(x,y) = -\frac{1}{2\lambda}\mathbb{E}_{\mathrm{P}} h\Phi\,.$$

This makes it clear that (5.8) can be viewed as a "continuous" and "quantified" version of the representer theorem.

The approach above yielded a quantified version of the representer theorem for *differentiable* losses. However, some important losses, such as the hinge loss and the pinball loss, are *not* differentiable, and since we are also interested in distributions P having point masses, this non-differentiability can cause problems even if it only occurs at one point. On the other hand, we are particularly interested in *convex* losses since these promise an efficient algorithmic treatment. Fortunately, for convex functions there exists a concept weaker than the derivative, for which the basic rules of calculus still hold. The following definition introduces this concept.

Definition 5.7. *Let E be a Banach space, $f : E \to \mathbb{R}\cup\{\infty\}$ be a convex function and $w \in E$ be an element with $f(w) \neq \infty$. Then the **subdifferential** of f at w is defined by*

$$\partial f(w) := \bigl\{w' \in E' : \langle w', v - w\rangle \leq f(v) - f(w) \text{ for all } v \in E\bigr\}\,.$$

Roughly speaking, the subdifferential $\partial f(w)$ contains all functionals describing the affine hyperplanes that are dominated by f and that are equal to it at w. For an illustration of this interpretation, see Figure 5.1. In particular, if f is Gâteaux differentiable at w, then $\partial f(w)$ contains only the derivative of f at w. This and some other important properties of the subdifferential can be found in Section A.6.2.

If $L : X \times Y \times \mathbb{R} \to [0, \infty)$ is a convex loss, we usually write $\partial L(x, y, t_0)$ for the subdifferential of the convex function $t \mapsto L(x, y, t)$ at the point $t_0 \in \mathbb{R}$. Moreover, we use analogous notation for supervised and unsupervised losses.

With these preparations, we can now state the main result of this section, which generalizes our considerations above to convex but not necessarily differentiable losses.

Theorem 5.8 (General representer theorem). *Let $p \in [1, \infty)$, P be a distribution on $X \times Y$, and $L : X \times Y \times \mathbb{R} \to [0, \infty)$ be a convex, P-integrable Nemitski loss of order p. Furthermore, let k be a bounded and measurable kernel on X with separable RKHS H and canonical feature map $\Phi : X \to H$. Then, for all $\lambda > 0$, there exists an $h \in \mathcal{L}_{p'}(\mathrm{P})$ such that*

$$h(x,y) \in \partial L\bigl(x, y, f_{\mathrm{P},\lambda}(x)\bigr)\,, \qquad (x,y) \in X \times Y, \tag{5.9}$$

$$f_{\mathrm{P},\lambda} = -\frac{1}{2\lambda}\, \mathbb{E}_{\mathrm{P}} h \Phi\,, \tag{5.10}$$

where p' is the conjugate exponent of p defined by $1/p' + 1/p = 1$.

Proof. Recall that L is a continuous loss since it is convex and finite. Moreover, L is a P-integrable Nemitski loss of order p, and hence we see by an almost literal repetition of the proof of Lemma 2.17 that $R : L_p(\mathrm{P}) \to [0, \infty)$ defined by

$$R(f) := \int_{X \times Y} L\bigl(x, y, f(x,y)\bigr)\, d\mathrm{P}(x,y)\,, \qquad f \in L_p(\mathrm{P}),$$

is well-defined and continuous.[1] Furthermore, Proposition A.6.13 shows that the subdifferential of R can be computed by

$$\partial R(f) = \bigl\{ h \in L_{p'}(\mathrm{P}) : h(x,y) \in \partial L(x, y, f(x,y)) \text{ for P-almost all } (x,y) \bigr\}.$$

Now, we easily infer from Lemma 4.23 that the inclusion map $I : H \to L_p(\mathrm{P})$ defined by $(If)(x,y) := f(x)$, $f \in H$, $(x,y) \in X \times Y$, is a bounded linear operator. Moreover, for $h \in L_{p'}(\mathrm{P})$ and $f \in H$, the reproducing property yields

$$\langle h, If \rangle_{L_{p'}(\mathrm{P}), L_p(\mathrm{P})} = \mathbb{E}_{\mathrm{P}} h If = \mathbb{E}_{\mathrm{P}} h \langle f, \Phi \rangle_H = \langle f, \mathbb{E}_{\mathrm{P}} h \Phi \rangle_H = \langle \iota \mathbb{E}_{\mathrm{P}} h \Phi, f \rangle_{H', H}\,,$$

where $\iota : H \to H'$ is the Fréchet-Riesz isomorphism described in Theorem A.5.12. Consequently, the adjoint operator I' of I is given by $I'h = \iota \mathbb{E}_{\mathrm{P}} h \Phi$,

[1] If we write $\bar{X} := X \times Y$, this can also be seen by interpreting R as the risk of the unsupervised, continuous loss $\bar{L} : \bar{X} \times \mathbb{R} \to [0, \infty)$, defined by $\bar{L}(\bar{x}, t) := L(x, y, t)$, $(x, y) := \bar{x} \in \bar{X}$, $t \in \mathbb{R}$.

$h \in L_{p'}(\mathrm{P})$. Moreover, the L-risk functional $\mathcal{R}_{L,\mathrm{P}} : H \to [0,\infty)$ restricted to H satisfies $\mathcal{R}_{L,\mathrm{P}} = R \circ I$, and hence the chain rule for subdifferentials (see Proposition A.6.12) yields $\partial\mathcal{R}_{L,\mathrm{P}}(f) = \partial(R \circ I)(f) = I'\partial R(If)$ for all $f \in H$. Applying the formula for $\partial R(f)$ thus yields

$$\begin{aligned}&\partial\mathcal{R}_{L,\mathrm{P}}(f)\\ &= \bigl\{\iota\mathbb{E}_{\mathrm{P}}h\Phi : h \in L_{p'}(\mathrm{P}) \text{ with } h(x,y) \in \partial L(x,y,f(x)) \text{ P-almost surely}\bigr\}\end{aligned}$$

for all $f \in H$. In addition, $f \mapsto \|f\|_H^2$ is Fréchet differentiable and its derivative at f is $2\iota f$ for all $f \in H$. By picking suitable representations of $h \in L_{p'}(\mathrm{P})$, Proposition A.6.12 thus gives

$$\begin{aligned}&\partial\mathcal{R}^{reg}_{L,\mathrm{P},\lambda}(f)\\ &= 2\lambda\iota f + \bigl\{\iota\mathbb{E}_{\mathrm{P}}h\Phi : h \in \mathcal{L}_{p'}(\mathrm{P}) \text{ with } h(x,y) \in \partial L(x,y,f(x)) \text{ for all } (x,y)\bigr\}\end{aligned}$$

for all $f \in H$. Now recall that $\mathcal{R}^{reg}_{L,\mathrm{P},\lambda}(\,\cdot\,)$ has a minimum at $f_{\mathrm{P},\lambda}$, and therefore we have $0 \in \partial\mathcal{R}^{reg}_{L,\mathrm{P},\lambda}(f_{\mathrm{P},\lambda})$ by another application of Proposition A.6.12. This together with the injectivity of ι yields the assertion. □

5.3 Stability of Infinite-Sample SVMs

Given a distribution P for which the general SVM solution $f_{\mathrm{P},\lambda}$ exists, one may ask how this solution changes if the underlying distribution P changes. The goal of this section is to answer this question with the help of the generalized representer theorem established in Section 5.2. The results we derive in this direction will be crucial for the stability-based statistical analysis in Sections 6.4 and 9.2. Moreover, it will be a key element in the robustness considerations of Sections 10.3 and 10.4.

Let us begin with the following theorem that, roughly speaking, provides the Lipschitz continuity of the map $\mathrm{P} \mapsto f_{\mathrm{P},\lambda}$.

Theorem 5.9. *Let $p \in [1,\infty)$ be a real number and $p' \in (1,\infty]$ be its conjugate defined by $\frac{1}{p} + \frac{1}{p'} = 1$. Furthermore, let $L : X \times Y \times \mathbb{R} \to [0,\infty)$ be a convex loss function and P be a distribution on $X \times Y$ such that L is a P-integrable Nemitski loss of order p. Furthermore, let k be a bounded and measurable kernel on X with separable RKHS H and canonical feature map $\Phi : X \to H$. Then, for all $\lambda > 0$, there exists an $h \in \mathcal{L}_{p'}(\mathrm{P})$ such that*

$$h(x,y) \in \partial L\bigl(x,y,f_{\mathrm{P},\lambda}(x)\bigr)\,, \qquad (x,y) \in X \times Y, \tag{5.11}$$

$$f_{\mathrm{P},\lambda} = -\frac{1}{2\lambda}\,\mathbb{E}_{\mathrm{P}}\,h\Phi\,, \tag{5.12}$$

$$h \in \mathcal{L}_1(\bar{\mathrm{P}})\,, \tag{5.13}$$

$$\bigl\|\, f_{\mathrm{P},\lambda} - f_{\bar{\mathrm{P}},\lambda}\,\bigr\|_H \le \frac{1}{\lambda}\bigl\|\,\mathbb{E}_{\mathrm{P}}h\Phi - \mathbb{E}_{\bar{\mathrm{P}}}h\Phi\,\bigr\|_H\,, \tag{5.14}$$

for all distributions $\bar{\mathrm{P}}$ on $X \times Y$ for which L is a $\bar{\mathrm{P}}$-integrable Nemitski loss.

Proof. By Theorem 5.8, there exists an $h \in \mathcal{L}_{p'}(\mathrm{P})$ satisfying (5.11) and (5.12). Let us first show that h is $\bar{\mathrm{P}}$-integrable, i.e., that (5.13) holds. To this end, observe that, since k is a bounded kernel, we have

$$\|f_{\mathrm{P},\lambda}\|_\infty \leq \|k\|_\infty \|f_{\mathrm{P},\lambda}\|_H \leq \|k\|_\infty \sqrt{\frac{\mathcal{R}_{L,\mathrm{P}}(0)}{\lambda}} =: B_\lambda < \infty \tag{5.15}$$

by Lemma 4.23 and (5.4). Moreover, since L is a $\bar{\mathrm{P}}$-integrable Nemitski loss, there exist a $\bar{b} \in \mathcal{L}_1(\bar{\mathrm{P}})$ and an increasing function $\bar{h} : [0,\infty) \to [0,\infty)$ with

$$L(x,y,t) \leq \bar{b}(x,y) + \bar{h}(|t|)\,, \qquad\qquad (x,y,t) \in X \times Y \times \mathbb{R}\,.$$

Now (5.11) and Proposition A.6.11 with $\delta := 1$ yield

$$|h(x,y)| \leq \sup\bigl|\partial L(x,y,f_{\mathrm{P},\lambda}(x))\bigr| \leq \bigl|L(x,y,\,\cdot\,)_{|[-1+\|f_{\mathrm{P},\lambda}\|_\infty, 1+\|f_{\mathrm{P},\lambda}\|_\infty]}\bigr|_1\,,$$

and hence Lemma A.6.5 together with (5.15) shows

$$\begin{aligned} |h(x,y)| \leq \bigl|L(x,y,\,\cdot\,)_{|[-1+B_\lambda,1+B_\lambda]}\bigr|_1 &\leq \frac{1}{1+B_\lambda}\,\bigl\|L(x,y,\,\cdot\,)_{|[-2+2B_\lambda,2+2B_\lambda]}\bigr\|_\infty \\ &\leq \frac{\bar{b}(x,y)+\bar{h}(2+2B_\lambda)}{1+B_\lambda} \end{aligned} \tag{5.16}$$

for all $(x,y) \in X \times Y$. From this we deduce $h \in \mathcal{L}_1(\bar{\mathrm{P}})$.

Let us now establish (5.14). To this end, observe that by (5.11) and the definition of the subdifferential, we have

$$h(x,y)\bigl(f_{\bar{\mathrm{P}},\lambda}(x) - f_{\mathrm{P},\lambda}(x)\bigr) \leq L\bigl(x,y,f_{\bar{\mathrm{P}},\lambda}(x)\bigr) - L\bigl(x,y,f_{\mathrm{P},\lambda}(x)\bigr)$$

for all $(x,y) \in X \times Y$. By integrating with respect to $\bar{\mathrm{P}}$, we hence obtain

$$\langle f_{\bar{\mathrm{P}},\lambda} - f_{\mathrm{P},\lambda}\,,\, \mathbb{E}_{\bar{\mathrm{P}}} h\Phi\rangle \;\leq\; \mathcal{R}_{L,\bar{\mathrm{P}}}(f_{\bar{\mathrm{P}},\lambda}) - \mathcal{R}_{L,\bar{\mathrm{P}}}(f_{\mathrm{P},\lambda})\,. \tag{5.17}$$

Moreover, an easy calculation shows

$$2\lambda\,\langle f_{\bar{\mathrm{P}},\lambda} - f_{\mathrm{P},\lambda}\,,\, f_{\mathrm{P},\lambda}\rangle + \lambda\|f_{\mathrm{P},\lambda} - f_{\bar{\mathrm{P}},\lambda}\|_H^2 = \lambda\|f_{\bar{\mathrm{P}},\lambda}\|_H^2 - \lambda\,\|f_{\mathrm{P},\lambda}\|_H^2\,. \tag{5.18}$$

By combining (5.17) and (5.18), we then find

$$\begin{aligned} \bigl\langle f_{\bar{\mathrm{P}},\lambda} - f_{\mathrm{P},\lambda}, \mathbb{E}_{\bar{\mathrm{P}}} h\Phi + 2\lambda f_{\mathrm{P},\lambda}\bigr\rangle + \lambda\|f_{\mathrm{P},\lambda} - f_{\bar{\mathrm{P}},\lambda}\|_H^2 &\leq \mathcal{R}^{reg}_{L,\bar{\mathrm{P}},\lambda}(f_{\bar{\mathrm{P}},\lambda}) - \mathcal{R}^{reg}_{L,\bar{\mathrm{P}},\lambda}(f_{\mathrm{P},\lambda}) \\ &\leq 0\,, \end{aligned}$$

and consequently the representation $f_{\mathrm{P},\lambda} = -\frac{1}{2\lambda}\,\mathbb{E}_{\mathrm{P}} h\Phi$ yields

$$\begin{aligned} \lambda\|f_{\mathrm{P},\lambda} - f_{\bar{\mathrm{P}},\lambda}\|_H^2 &\leq \bigl\langle f_{\mathrm{P},\lambda} - f_{\bar{\mathrm{P}},\lambda}\,,\, \mathbb{E}_{\bar{\mathrm{P}}} h\Phi - \mathbb{E}_{\mathrm{P}} h\Phi\bigr\rangle \\ &\leq \|f_{\mathrm{P},\lambda} - f_{\bar{\mathrm{P}},\lambda}\|_H \,\cdot\, \bigl\|\mathbb{E}_{\bar{\mathrm{P}}} h\Phi - \mathbb{E}_{\mathrm{P}} h\Phi\bigr\|_H\,. \end{aligned}$$

From this we easily obtain (5.14). □

For the applications in later chapters, it is often necessary to have more information on the representing function h in (5.12). For two important types of losses, such additional information is presented in the following two corollaries.

Corollary 5.10. *Let $L : X \times Y \times \mathbb{R} \to [0, \infty)$ be a convex, locally Lipschitz continuous loss and P be a distribution on $X \times Y$ with $\mathcal{R}_{L,\mathrm{P}}(0) < \infty$. Furthermore, let k be a bounded and measurable kernel on X with separable RKHS H and canonical feature map $\Phi : X \to H$. We write*

$$B_\lambda := \|k\|_\infty \left(\frac{\mathcal{R}_{L,\mathrm{P}}(0)}{\lambda} \right)^{1/2} , \qquad\qquad \lambda > 0.$$

Then, for all $\lambda > 0$, there exists a bounded measurable function $h : X \times Y \to \mathbb{R}$ such that, for all distributions $\bar{\mathrm{P}}$ on $X \times Y$ with $\mathcal{R}_{L,\bar{\mathrm{P}}}(0) < \infty$, we have

$$\begin{aligned} h(x,y) &\in \partial L\bigl(x, y, f_{\mathrm{P},\lambda}(x)\bigr) , &\qquad (x,y) \in X \times Y, \qquad (5.19)\\ f_{\mathrm{P},\lambda} &= -\frac{1}{2\lambda}\, \mathbb{E}_{\mathrm{P}}\, h\Phi , & (5.20)\\ \|h\|_\infty &\le |L|_{B_\lambda,1} , & (5.21)\\ \bigl\| f_{\mathrm{P},\lambda} - f_{\bar{\mathrm{P}},\lambda} \bigr\|_H &\le \frac{1}{\lambda} \bigl\| \mathbb{E}_{\mathrm{P}} h\Phi - \mathbb{E}_{\bar{\mathrm{P}}} h\Phi \bigr\|_H . & (5.22) \end{aligned}$$

Proof. Let us fix a $\lambda > 0$ and write $B := B_\lambda$. By Lemma 4.23 and Corollary 5.3, we then know that

$$\|f_{\mathrm{P},\lambda}\|_\infty \;\le\; \|k\|_\infty \, \|f_{\mathrm{P},\lambda}\|_H \;\le\; \|k\|_\infty \left(\frac{\mathcal{R}_{L,\mathrm{P}}(0)}{\lambda} \right)^{1/2} = B .$$

For $(x, y) \in X \times Y$, we further know that $L(x, y, \cdot\,) : \mathbb{R} \to [0, \infty)$ is convex, and hence Lemma A.6.16 shows that there exists a convex and Lipschitz continuous function $\tilde{L}(x, y, \cdot\,) : \mathbb{R} \to [0, \infty)$ with

$$\begin{aligned} \tilde{L}(x, y, \cdot\,)_{|[-B,B]} &= L(x, y, \cdot\,)_{|[-B,B]} \\ |\tilde{L}(x, y, \cdot\,)|_1 &= \bigl| L(x, y, \cdot\,)_{|[-B,B]} \bigr|_1 \\ \partial \tilde{L}(x, y, t) &\subset \partial L(x, y, t) , \qquad\qquad t \in [-B, B]. \end{aligned}$$

Moreover, considering the proof of Lemma A.6.16, we see that $\tilde{L} : X \times Y \times \mathbb{R} \to [0, \infty)$ can also be assumed to be measurable. Therefore, $\tilde{L}$ is a convex, Lipschitz continuous loss function with $|\tilde{L}|_1 = |L|_{B,1}$ and $\mathcal{R}_{\tilde{L},\mathrm{P}}(0) = \mathcal{R}_{L,\mathrm{P}}(0)$. Consequently, Corollary 5.3 shows that the general SVM solution

$$\tilde{f}_{\mathrm{P},\lambda} \in \arg\min_{f \in H} \lambda \|f\|_H^2 + \mathcal{R}_{\tilde{L},\mathrm{P}}(f)$$

exists and satisfies $\|\tilde{f}_{\mathrm{P},\lambda}\|_\infty \le B$. Furthermore, we have $\mathcal{R}_{\tilde{L},\mathrm{P}}(f) = \mathcal{R}_{L,\mathrm{P}}(f)$ for all measurable $f : X \to [-B, B]$, and hence $\tilde{f}_{\mathrm{P},\lambda}$ is also a minimizer of

$\mathcal{R}^{reg}_{L,\mathrm{P},\lambda}(\,\cdot\,)$. By the uniqueness of $f_{\mathrm{P},\lambda}$, we thus find $f_{\mathrm{P},\lambda} = \tilde{f}_{\mathrm{P},\lambda}$. Now recall that the Lipschitz continuity of $\tilde{L}$ gives

$$\tilde{L}(x,y,t) \leq L(x,y,0) + |\tilde{L}|_1\,|t|\,, \qquad\qquad (x,y) \in X \times Y,\ t \in \mathbb{R},$$

i.e., $\tilde{L}$ is a Nemitski loss of order $p := 1$. Therefore, Theorem 5.8 gives a bounded measurable function $h : X \times Y \to \mathbb{R}$ such that

$$h(x,y) \in \partial\tilde{L}\bigl(x,y,\tilde{f}_{\mathrm{P},\lambda}(x)\bigr) \subset \partial L\bigl(x,y,f_{\mathrm{P},\lambda}(x)\bigr), \qquad (x,y) \in X \times Y, \quad (5.23)$$

and $-\frac{1}{2\lambda}\,\mathbb{E}_{\mathrm{P}}h\Phi = \tilde{f}_{\mathrm{P},\lambda} = f_{\mathrm{P},\lambda}$. In other words, we have shown (5.19) and (5.20). In addition, combining (5.23) with Proposition A.6.11 yields

$$\bigl|h(x,y)\bigr| \;\leq\; \sup\bigl\{|t| : t \in \partial\tilde{L}(x,y,\tilde{f}_{\mathrm{P},\lambda}(x))\bigr\} \;\leq\; |\tilde{L}|_1 \;=\; |L|_{B,1}$$

for all $(x,y) \in X \times Y$, i.e., we have shown (5.21). Finally, (5.22) can be shown as in the proof of Theorem 5.9. □

Recall that convex, margin-based losses are locally Lipschitz continuous, and hence Corollary 5.10 provides a generalized representer theorem together with the Lipschitz continuity of $\mathrm{P} \mapsto f_{\mathrm{P},\lambda}$ for this important class of loss functions. Let us now consider distance-based losses.

Corollary 5.11. *Let $L : \mathbb{R} \times \mathbb{R} \to [0,\infty)$ be a convex, distance-based loss of upper growth type $p \geq 1$ with representing function $\psi : \mathbb{R} \to [0,\infty)$. Furthermore, let P be a distribution on $X \times \mathbb{R}$ with $|\mathrm{P}|_p < \infty$ and k be a bounded and measurable kernel on X with separable RKHS H and canonical feature map $\Phi : X \to H$. Then there exists a constant $c_L > 0$ depending only on L such that for all $\lambda > 0$ there exists a measurable function $h : X \times Y \to \mathbb{R}$ such that*

$$h(x,y) \in -\partial\psi\bigl(y - f_{\mathrm{P},\lambda}(x)\bigr)\,, \qquad\qquad (x,y) \in X \times Y, \qquad (5.24)$$

$$f_{\mathrm{P},\lambda} = -\frac{1}{2\lambda}\,\mathbb{E}_{\mathrm{P}}\,h\Phi\,, \qquad (5.25)$$

$$\|h\|_{L_s(\bar{\mathrm{P}})} \leq 8^p c_L \left(1 + |\bar{\mathrm{P}}|_q^{p-1} + \|f_{\mathrm{P},\lambda}\|_\infty^{p-1}\right), \qquad (5.26)$$

$$\bigl\|\, f_{\mathrm{P},\lambda} - f_{\bar{\mathrm{P}},\lambda} \,\bigr\|_H \leq \frac{1}{\lambda}\bigl\|\, \mathbb{E}_{\mathrm{P}}h\Phi - \mathbb{E}_{\bar{\mathrm{P}}}h\Phi \,\bigr\|_H\,, \qquad (5.27)$$

for all $q \in [p,\infty]$, all distributions $\bar{\mathrm{P}}$ on $X \times Y$ with $|\bar{\mathrm{P}}|_q < \infty$, and $s := \frac{q}{p-1}$.

Proof. Recall that L is a P-integrable Nemitski loss of order p by Lemma 2.38, and by Theorem 5.9 there thus exists an $h \in \mathcal{L}_{p'}(\mathrm{P})$ satisfying (5.11), (5.12), and (5.14). Since (5.12) equals (5.25) and (5.14) equals (5.27), it remains to show (5.24) and (5.26). To this end, recall that ψ satisfies $L(y,t) = \psi(y - t)$, $y,t \in \mathbb{R}$, and hence it is convex. By Proposition A.6.12, we then have $\partial L(y,t) = -\partial\psi(y-t)$ for $y,t \in \mathbb{R}$, and hence (5.11) implies (5.24). Moreover, for $p = 1$, the loss L is Lipschitz continuous by Lemma 2.36, and consequently

(5.26) follows from Proposition A.6.11. Therefore let us now consider the case $p > 1$. For $(x, y) \in X \times Y$ with $r := |y - f_{\mathrm{P},\lambda}(x)| \geq 1$, Proposition A.6.11 with $\delta := r$ and Lemma 2.36 then yield

$$|h(x,y)| \leq |\psi_{|[-2r,2r]}|_1 \leq r^{-1} \, \|\psi_{|[4r,4r]}\|_\infty \leq cr^{-1} \left((4r)^p + 1\right) \leq c\, 4^{p+1} \, r^{p-1} \, ,$$

where $c > 0$ is the constant arising in the upper growth type definition. Moreover, for $(x, y) \in X \times Y$ with $r := |y - f_{\mathrm{P},\lambda}(x)| \leq 1$, Proposition A.6.11 gives

$$|h(x,y)| \leq |\psi_{|[-2r,2r]}|_1 \leq |\psi_{|[-2,2]}|_1 \, .$$

Together, these estimates show that for all $(x, y) \in X \times Y$ we have

$$|h(x,y)| \leq 4^p \, c_L \max\bigl\{1, |y - f_{\mathrm{P},\lambda}(x)|^{p-1}\bigr\} \, , \tag{5.28}$$

where c_L is a suitable constant depending only on the loss function L. For $q = \infty$, we then easily find the assertion, and hence let us assume $q \in [p, \infty)$. In this case, the inequality above yields

$$|h(x,y)|^s \leq 4^{ps} c_L^s \max\bigl\{1, |y - f_{\mathrm{P},\lambda}(x)|^q\bigr\} \leq 4^{ps} 2^{q-1} c_L^s \bigl(1 + |y|^q + |f_{\mathrm{P},\lambda}(x)|^q\bigr) \, ,$$

and using $4^p 2^{\frac{q-1}{s}} \leq 8^p$ we consequently find

$$\|h\|_{L_s(\bar{\mathrm{P}})} \leq 8^p c_L \bigl(1 + |\bar{\mathrm{P}}|_q^{p-1} + \|f_{\mathrm{P},\lambda}\|_\infty^{p-1}\bigr) \, . \qquad \square$$

Note that the larger we can choose the real number q in the preceding corollary, the larger the corresponding s becomes, i.e., the stronger the integrability condition on h becomes. In particular, the case $q = \infty$ yields $s = \infty$, i.e., the representing function h is bounded.

The following corollary shows that Theorem 5.9 holds under weaker assumptions if all distributions involved are empirical.

Corollary 5.12. *Let $L : X \times Y \times \mathbb{R} \to [0, \infty)$ be a convex loss function, $n \geq 1$, and $D \in (X \times Y)^n$. Furthermore, let H be the RKHS of a kernel on X and $\Phi : X \to H$ be its canonical feature map. Moreover, let $\lambda > 0$ and*

$$B_\lambda := \|k\|_\infty \left(\frac{\mathcal{R}_{L,\mathrm{D}}(0)}{\lambda}\right)^{1/2} . \tag{5.29}$$

Then there exists a function $h : X \times Y \to \mathbb{R}$ such that, for all $m \geq 1$ and all $\bar{D} \in (X \times Y)^m$, we have

$$\begin{aligned} h(x,y) &\in \partial L\bigl(x, y, f_{\mathrm{D},\lambda}(x)\bigr) \, , \qquad\qquad (x,y) \in D \cup \bar{D}, \\ f_{\mathrm{D},\lambda} &= -\frac{1}{2\lambda} \, \mathbb{E}_{\mathrm{D}} \, h\Phi \, , \\ \|h\|_\infty &\leq |L|_{B_\lambda,1} \, , \\ \|f_{\mathrm{D},\lambda} - f_{\bar{\mathrm{D}},\lambda}\|_H &\leq \frac{1}{\lambda} \bigl\| \, \mathbb{E}_{\mathrm{D}} h\Phi - \mathbb{E}_{\bar{\mathrm{D}}} h\Phi \, \bigr\|_H \, . \end{aligned}$$

Proof. We write $D = ((x_1, y_1), \dots, (x_n, y_n))$ and $\bar{D} = ((\bar{x}_1, \bar{y}_1), \dots, (\bar{x}_m, \bar{y}_m))$. Furthermore, we define

$$\begin{aligned} X' &:= \{x_i : i = 1, \dots, n\} \cup \{\bar{x}_i : i = 1, \dots, m\}\,, \\ Y' &:= \{y_i : i = 1, \dots, n\} \cup \{\bar{y}_i : i = 1, \dots, m\}\,, \end{aligned}$$

and equip both sets with the discrete σ-algebra. Then $k_{|X' \times X'}$ is a bounded measurable kernel with finite-dimensional, and hence separable, RKHS. In addition, both D and $\bar{\mathrm{D}}$ are distributions on $X' \times Y'$. Furthermore, $L(x, y, \cdot\,) : \mathbb{R} \to [0, \infty)$ is convex for all $(x, y) \in X' \times Y'$ and hence L is locally Lipschitz continuous. Since X' and Y' are finite sets, we then see that L restricted to $X' \times Y'$ is locally Lipschitz. Moreover, we obviously have both $\mathcal{R}_{L,\mathrm{D}}(0) < \infty$ and $\mathcal{R}_{L,\bar{\mathrm{D}}}(0) < \infty$, and hence the assertion follows from Corollary 5.10. □

Let us finally use the results above to show that, under some additional conditions, the map $D \mapsto f_{\mathrm{D},\lambda}$ is continuous.

Lemma 5.13. *Let X be a metric space and $L : X \times Y \times \mathbb{R} \to [0, \infty)$ be a continuous function that is a convex and differentiable loss. Furthermore, let H be the RKHS of a continuous kernel on X, $n \geq 1$, and $\lambda > 0$. Then the map $(X \times Y)^n \to H$ defined by $D \mapsto f_{\mathrm{D},\lambda}$ is continuous.*

Proof. Let us fix two sample sets $D := ((x_1, y_1), \dots, (x_n, y_n)) \in (X \times Y)^n$ and $\bar{D} := ((\bar{x}_1, \bar{y}_1), \dots, (\bar{x}_m, \bar{y}_m)) \in (X \times Y)^n$. By Corollary 5.12, we then obtain

$$\| f_{\mathrm{D},\lambda} - f_{\bar{\mathrm{D}},\lambda} \|_H \leq \frac{1}{\lambda n} \left\| \sum_{i=1}^{n} L'\big(x_i, y_i, f_{\mathrm{D},\lambda}(x_i)\big) \Phi(x_i) - L'\big(\bar{x}_i, \bar{y}_i, f_{\mathrm{D},\lambda}(\bar{x}_i)\big) \Phi(\bar{x}_i) \right\|.$$

Now Lemma 4.29 shows that both $f_{\mathrm{D},\lambda}$ and $\Phi : X \to H$ are continuous. Moreover, every convex differentiable function is continuously differentiable by Proposition A.6.14, and hence we obtain the assertion. □

5.4 Behavior for Small Regularization Parameters

In this section, we investigate how the general SVM solution $f_{\mathrm{P},\lambda}$ and its associated risk $\mathcal{R}_{L,\mathrm{P}}(f_{\mathrm{P},\lambda})$ behaves for vanishing regularization parameter, i.e., for $\lambda \to 0$. In addition, we compare the behavior of the minimized regularized risk with the approximation error of the scaled unit balls $\lambda^{-1} B_H$.

Let us begin by introducing a new quantity. To this end, let $L : X \times Y \times \mathbb{R} \to [0, \infty)$ be a loss, k be a measurable kernel over X with RKHS H, and P be a distribution on $X \times Y$. Then we write

$$\mathcal{R}^*_{L,\mathrm{P},H} := \inf_{f \in H} \mathcal{R}_{L,\mathrm{P}}(f) \tag{5.30}$$

for the smallest possible L-risk on H. Moreover, we say that an element $f^* \in H$ **minimizes the L-risk in H** if it satisfies $\mathcal{R}_{L,\mathrm{P}}(f^*) = \mathcal{R}^*_{L,\mathrm{P},H}$. Finally, we need the following fundamental definition, which is closely related to the minimization of the regularized risk (5.2).

Definition 5.14. *Let $L : X \times Y \times \mathbb{R} \to [0, \infty)$ be a loss, H be the RKHS of a measurable kernel on X, and P be a distribution on $X \times Y$ with $\mathcal{R}^*_{L,\mathrm{P},H} < \infty$. Then we define the* ***approximation error function*** *$A_2 : [0, \infty) \to [0, \infty)$ by*

$$A_2(\lambda) := \inf_{f\in H} \lambda\|f\|_H^2 + \mathcal{R}_{L,\mathrm{P}}(f) - \mathcal{R}^*_{L,\mathrm{P},H} \,, \qquad \lambda \geq 0.$$

Our first lemma collects some simple properties of the function A_2.

Lemma 5.15 (Properties of the approximation error function). *Let $L : X\times Y\times\mathbb{R} \to [0, \infty)$ be a loss, H be the RKHS of a measurable kernel on X, and P be a distribution on $X\times Y$ with $\mathcal{R}^*_{L,\mathrm{P},H} < \infty$. Then $A_2 : [0, \infty) \to [0, \infty)$ is increasing, concave, and continuous. Moreover, we have $A_2(0) = 0$ and*

$$\begin{aligned} \frac{A_2(\kappa)}{\kappa} &\leq \frac{A_2(\lambda)}{\lambda}\,, & 0 < \lambda \leq \kappa, \qquad (5.31)\\ A_2(\lambda) &\leq \mathcal{R}_{L,\mathrm{P}}(0) - \mathcal{R}^*_{L,\mathrm{P},H}\,, & \lambda \geq 0. \end{aligned}$$

In addition, $A_2(\,\cdot\,)$ is subadditive in the sense of

$$A_2(\lambda + \kappa) \;\leq\; A_2(\lambda) + A_2(\kappa)\,, \qquad \lambda, \kappa \geq 0.$$

Finally, if there exists a function $h : [0, 1] \to [0, \infty)$ with $\lim_{\lambda\to 0^+} h(\lambda) = 0$ and $A_2(\lambda) \leq \lambda h(\lambda)$ for all $\lambda \in [0, 1]$, then we have $A_2(\lambda) = 0$ for all $\lambda \geq 0$, and 0 minimizes $\mathcal{R}_{L,\mathrm{P}}$ in H.

Proof. The definition of the approximation error function immediately gives $A_2(0) = 0$. Moreover, $A_2(\,\cdot\,)$ is an infimum over a family of affine linear and increasing functions, and hence we see that A_2 is concave, continuous, and increasing by Lemma A.6.4. Furthermore, (5.31) follows from the concavity, namely

$$A_2(\lambda) = A_2\Big(\frac{\lambda}{\kappa}\kappa + \Big(1 - \frac{\lambda}{\kappa}\Big)0\Big) \geq \frac{\lambda}{\kappa}A_2(\kappa) + \Big(1 - \frac{\lambda}{\kappa}\Big)A_2(0) = \frac{\lambda}{\kappa}A_2(\kappa)\,.$$

In addition, for $\lambda \geq 0$, we obtain from the definition of A_2 that

$$A_2(\lambda) \leq \lambda\|0\|_H^2 + \mathcal{R}_{L,\mathrm{P}}(0) - \mathcal{R}^*_{L,\mathrm{P},H} = \mathcal{R}_{L,\mathrm{P}}(0) - \mathcal{R}^*_{L,\mathrm{P},H}\,.$$

In order to show the subadditivity, it suffices to consider $\lambda, \kappa > 0$. Without loss of generality, we may additionally assume $\lambda \leq \kappa$. Using the already proved (5.31) twice, we then obtain

$$A_2(\lambda + \kappa) \leq (\lambda + \kappa)\kappa^{-1}A_2(\kappa) = \lambda\kappa^{-1}A_2(\kappa) + A_2(\kappa) \leq A_2(\lambda) + A_2(\kappa)\,.$$

Finally, let us assume that we have $A_2(\lambda) \leq \lambda h(\lambda)$ for all $\lambda \in [0, 1]$. Using our previous results, we then obtain

$$A_2(1) \;\leq\; \lambda^{-1}A_2(\lambda) \;\leq\; h(\lambda)\,, \qquad \lambda \in (0, 1].$$

For $\lambda \to 0$, we then find $A_2(1) = 0$, and since A_2 is a non-negative and concave function with $A_2(0) = 0$, we then find $A_2(\lambda) = 0$ for all $\lambda \geq 0$. The last assertion is a trivial consequence of the latter. □

Now assume for a moment that the general SVM solution $f_{\mathrm{P},\lambda}$ exists for all $\lambda > 0$. The continuity of A_2 at 0 together with $A_2(0) = 0$ then shows both

$$\lim_{\lambda \to 0} \lambda \|f_{\mathrm{P},\lambda}\|_H^2 + \mathcal{R}_{L,\mathrm{P}}(f_{\mathrm{P},\lambda}) = \mathcal{R}^*_{L,\mathrm{P},H} \tag{5.32}$$

and $\lim_{\lambda \to 0} \mathcal{R}_{L,\mathrm{P}}(f_{\mathrm{P},\lambda}) = \mathcal{R}^*_{L,\mathrm{P},H}$. In other words, the (regularized) risk of the SVM solution $f_{\mathrm{P},\lambda}$ tends to the minimal risk in H for vanishing regularization term λ. Since this suggests that the behavior of A_2 becomes particularly interesting for $\lambda \to 0$, we will mainly focus on this limit behavior. To this end, we need the following preparatory lemma.

Lemma 5.16. *Let $L : X \times Y \times \mathbb{R} \to [0, \infty)$ be a convex loss, H be the RKHS of a measurable kernel on X, and* P *be a distribution on $X \times Y$ with $\mathcal{R}^*_{L,\mathrm{P},H} < \infty$. Assume that there exists an $f_0^* \in H$ with $\mathcal{R}_{L,\mathrm{P}}(f_0^*) = \mathcal{R}^*_{L,\mathrm{P},H}$. Then there exists exactly one element $f^*_{L,\mathrm{P},H} \in H$ such that both*

$$\mathcal{R}_{L,\mathrm{P}}(f^*_{L,\mathrm{P},H}) = \mathcal{R}^*_{L,\mathrm{P},H}$$

and $\|f^*_{L,\mathrm{P},H}\| \leq \|f^*\|$ for all $f^* \in H$ satisfying $\mathcal{R}_{L,\mathrm{P}}(f^*) = \mathcal{R}^*_{L,\mathrm{P},H}$.

Proof. Let M denote the set of all elements f^* minimizing $\mathcal{R}_{L,\mathrm{P}}$ in H, i.e.

$$M := \left\{ f^* \in H : \mathcal{R}_{L,\mathrm{P}}(f^*) = \mathcal{R}^*_{L,\mathrm{P},H} \right\}.$$

Uniqueness. Assume that we have two different elements $f_1^*, f_2^* \in M$ satisfying $\|f_1^*\| \leq \|f\|$ and $\|f_2^*\| \leq \|f\|$ for all $f \in M$. Then we obviously have $\|f_1^*\| = \|f_2^*\|$. Moreover, the convexity of $\mathcal{R}_{L,\mathrm{P}}$ implies $\frac{1}{2}(f_1^* + f_2^*) \in M$, and the last statement in Lemma A.5.9 shows

$$\left\| \frac{1}{2}(f_1^* + f_2^*) \right\|_H^2 < \frac{\|f_1^*\|_H^2 + \|f_2^*\|_H^2}{2} = \|f_1^*\|_H^2 . \tag{5.33}$$

Since this contradicts our assumptions on f_1^*, there is at most one $f^*_{L,\mathrm{P},H}$.

Existence. Since L is a convex loss, it is continuous, and consequently the risk functional $\mathcal{R}_{L,\mathrm{P}} : H \to [0, \infty]$ is lower semi-continuous by Lemma 2.15. This shows that the set M of minimizing elements is closed. Using Lemma A.2.8, we then see that the map $N : H \to [0, \infty]$ defined by

$$N(f) := \begin{cases} \|f\| & \text{if } f \in M \\ \infty & \text{otherwise} \end{cases}$$

is lower semi-continuous. In addition, M is convex by the convexity of L, and hence so is N. Now observe that the set $\{f \in H : N(f) \leq \|f_0^*\|\}$ is non-empty and bounded, and consequently N has a global minimum at some $f_0 \in H$ by Theorem A.6.9. Obviously, this f_0 is the desired $f^*_{L,\mathrm{P},H}$. □

Let us now assume for a moment that we are in the situation of Lemma 5.16. If $f_{\mathrm{P},\lambda}$ exists for some $\lambda > 0$, it necessarily satisfies

$$\|f_{\mathrm{P},\lambda}\| \;\leq\; \|f^*_{L,\mathrm{P},H}\| \tag{5.34}$$

since otherwise we would find a contradiction by

$$\lambda\|f^*_{L,\mathrm{P},H}\|^2_H + \mathcal{R}_{L,\mathrm{P}}(f^*_{L,\mathrm{P},H}) \;<\; \lambda\|f_{\mathrm{P},\lambda}\|^2_H + \mathcal{R}_{L,\mathrm{P}}(f_{\mathrm{P},\lambda})\,.$$

Moreover, observe that for $\lambda = 0$ a simple calculation shows

$$\lambda\|f^*_{L,\mathrm{P},H}\|^2_H + \mathcal{R}_{L,\mathrm{P}}(f^*_{L,\mathrm{P},H}) = \inf_{f\in H} \lambda\|f\|^2_H + \mathcal{R}_{L,\mathrm{P}}(f)\,,$$

and hence we write $f_{\mathrm{P},0} := f^*_{L,\mathrm{P},H}$. With this notation, we can now formulate our first main result describing the behavior of the function $\lambda \mapsto f_{\mathrm{P},\lambda}$.

Theorem 5.17 (Continuity in the regularization parameter). *Let* P *be a distribution on* $X \times Y$, $L : X \times Y \times \mathbb{R} \to [0,\infty)$ *be a convex* P*-integrable Nemitski loss, and* H *be the RKHS of a bounded measurable kernel on* X*. Furthermore, let* $(\lambda_n) \subset (0,\infty)$ *be a sequence that converges to a real number* $\lambda \in [0,\infty)$*. If the sequence* $(f_{\mathrm{P},\lambda_n})$ *is bounded, then* $f_{\mathrm{P},\lambda}$ *exists and we have*

$$\lim_{n\to\infty} \|f_{\mathrm{P},\lambda_n} - f_{\mathrm{P},\lambda}\|_H = 0\,.$$

Proof. Since $(f_{\mathrm{P},\lambda_n})$ is bounded, Theorem A.5.6 shows that there exist an $f^* \in H$ and a subsequence $(f_{\mathrm{P},\lambda_{n_i}})$ such that

$$f_{\mathrm{P},\lambda_{n_i}} \to f^* \qquad \text{with respect to the weak topology in } H.$$

Moreover, since this subsequence is bounded, we may additionally assume that there exists a $c \geq 0$ such that $\|f_{\mathrm{P},\lambda_{n_i}}\| \to c$. Now recall that the Dirac functionals are contained in the dual H', and therefore weak convergence in H implies pointwise convergence. Lemma 2.17 thus shows $\mathcal{R}_{L,\mathrm{P}}(f_{\mathrm{P},\lambda_{n_i}}) \to \mathcal{R}_{L,\mathrm{P}}(f^*)$. Furthermore, by (A.21), the weak convergence of $(f_{\mathrm{P},\lambda_{n_i}})$ implies

$$\|f^*\| \leq \liminf_{i\to\infty} \|f_{\mathrm{P},\lambda_{n_i}}\| = \lim_{i\to\infty} \|f_{\mathrm{P},\lambda_{n_i}}\| = c\,, \tag{5.35}$$

and hence the continuity of $A_2(\,\cdot\,)$ established in Lemma 5.15 yields

$$\begin{aligned}
\lambda\|f^*\|^2 + \mathcal{R}_{L,\mathrm{P}}(f^*) - \mathcal{R}^*_{L,\mathrm{P},H} &\leq \lambda c^2 + \mathcal{R}_{L,\mathrm{P}}(f^*) - \mathcal{R}^*_{L,\mathrm{P},H} \\
&= \lim_{i\to\infty} \big(\lambda_{n_i}\|f_{\mathrm{P},\lambda_{n_i}}\|^2 + \mathcal{R}_{L,\mathrm{P}}(f_{\mathrm{P},\lambda_{n_i}}) - \mathcal{R}^*_{L,\mathrm{P},H}\big) \\
&= \lim_{i\to\infty} A_2(\lambda_{n_i}) \\
&= A_2(\lambda)\,.
\end{aligned} \tag{5.36}$$

Consequently, f^* minimizes the regularized L-risk. If $\lambda > 0$, this implies $f^* = f_{\mathrm{P},\lambda}$ by Lemma 5.1. Furthermore, if $\lambda = 0$, this means that f^* minimizes

$\mathcal{R}_{L,\mathrm{P}}$ in H, and by combining (5.35) with (5.34), we thus find $\|f^*\| \leq \|f^*_{L,\mathrm{P},H}\|$, i.e., we have $f^* = f^*_{L,\mathrm{P},H} = f_{\mathrm{P},0}$ by Lemma 5.16.

Now, using the equality $f^* = f_{\mathrm{P},\lambda}$ in (5.36), we obtain

$$\lambda \|f_{\mathrm{P},\lambda}\|^2 + \mathcal{R}_{L,\mathrm{P}}(f_{\mathrm{P},\lambda}) - \mathcal{R}^*_{L,\mathrm{P},H} \;=\; \lambda c^2 + \mathcal{R}_{L,\mathrm{P}}(f_{\mathrm{P},\lambda}) - \mathcal{R}^*_{L,\mathrm{P},H}\,,$$

i.e., we have $c = \|f_{\mathrm{P},\lambda}\|$. Therefore, we have both $f_{\mathrm{P},\lambda_{n_i}} \to f_{\mathrm{P},\lambda}$ weakly and $\|f_{\mathrm{P},\lambda_{n_i}}\| \to \|f_{\mathrm{P},\lambda}\|$. Together these convergences imply

$$\lim_{i\to\infty} \|f_{\mathrm{P},\lambda_{n_i}} - f_{\mathrm{P},\lambda}\|^2 = \lim_{i\to\infty} \bigl(\|f_{\mathrm{P},\lambda_{n_i}}\|^2 - 2\langle f_{\mathrm{P},\lambda_{n_i}}, f_{\mathrm{P},\lambda}\rangle + \|f_{\mathrm{P},\lambda}\|^2\bigr) = 0\,,$$

i.e., the subsequence $(f_{\mathrm{P},\lambda_{n_i}})$ converges to $f_{\mathrm{P},\lambda}$ with respect to $\|\cdot\|_H$. Finally, let us assume that (f_{P,λ_n}) does *not* converge to $f_{\mathrm{P},\lambda}$ in norm. Then there exists a $\delta > 0$ and a subsequence $(f_{\mathrm{P},\lambda_{n_j}})$ with $\|f_{\mathrm{P},\lambda_{n_j}} - f_{\mathrm{P},\lambda}\| > \delta$. However, applying the reasoning above to this subsequence gives a sub-subsequence converging to $f_{\mathrm{P},\lambda}$ and hence we find a contradiction. □

The preceding theorem has some interesting consequences on the behavior of both $\lambda \mapsto f_{\mathrm{P},\lambda}$ and $\lambda \mapsto A_2(\lambda)$. Let us begin with a result that characterizes the existence of $f^*_{L,\mathrm{P},H}$ in terms of $\lambda \mapsto \|f_{\mathrm{P},\lambda}\|$ and A_2.

Corollary 5.18. *Let* P *be a distribution on* $X \times Y$, $L : X \times Y \times \mathbb{R} \to [0,\infty)$ *be a convex* P*-integrable Nemitski loss, and* H *be the RKHS of a bounded measurable kernel on* X*. Then the following statements are equivalent:*

i) There exists an $f^* \in H$ *minimizing* $\mathcal{R}_{L,\mathrm{P}}$ *in* H*.*
ii) The function $\lambda \mapsto \|f_{\mathrm{P},\lambda}\|$ *is bounded on* $(0,\infty)$*.*
iii) There exists a constant $c > 0$ *with* $A_2(\lambda) \leq c\lambda$ *for all* $\lambda \geq 0$*.*

In addition, if one of the statements is satisfied, we have $A_2(\lambda) \;\leq\; \lambda\,\|f^*_{L,\mathrm{P},H}\|^2_H$ *for all* $\lambda \geq 0$ *and*

$$\lim_{\lambda\to 0^+} \|f_{\mathrm{P},\lambda} - f^*_{L,\mathrm{P},H}\|_H = 0\,.$$

Proof. ii) $\Rightarrow$ *i)*. Let $(\lambda_n) \subset (0,\infty)$ be a sequence with $\lambda_n \to 0$. Our assumption then guarantees that the sequence $(f_{\mathrm{P},\lambda_n})$ is bounded, and hence Theorem 5.17 shows that $f_{\mathrm{P},0} = f^*_{L,\mathrm{P},H}$ exists and that we have $f_{\mathrm{P},\lambda_n} \to f^*_{L,\mathrm{P},H}$.

i) $\Rightarrow$ *iii)*. By Lemma 5.16, we know that $f^*_{L,\mathrm{P},H}$ exists. Now the implication follows from the estimate

$$A_2(\lambda) \leq \lambda \|f^*_{L,\mathrm{P},H}\|^2 + \mathcal{R}_{L,\mathrm{P}}(f^*_{L,\mathrm{P},H}) - \mathcal{R}^*_{L,\mathrm{P},H} = \lambda \|f^*_{L,\mathrm{P},H}\|^2\,, \qquad \lambda \geq 0.$$

iii) $\Rightarrow$ *ii)*. This implication follows from the estimate

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

Let us now assume for a moment that L is the hinge loss and that H is the RKHS of a universal (or just a strictly positive definite) kernel. For a training set D without contradicting samples, i.e., $x_i = x_j$ implies $y_i = y_j$, it is then straightforward to see that there is an $f^* \in H$ with $\mathcal{R}_{L,\mathrm{D}}(f^*) = 0$. Obviously, this f^* minimizes $\mathcal{R}_{L,\mathrm{D}}$ in H and thus $f^*_{L,\mathrm{D},H} \in H$ does exist. Moreover, since $f^*_{L,\mathrm{D},H}$ has zero error on D and minimal norm among such minimizers, it coincides with the *hard margin* SVM solution. Consequently, Corollary 5.18 shows that the soft margin SVM solutions $f_{\mathrm{D},\lambda}$ converge to the hard margin solution if D is *fixed* and $\lambda \to 0$. Since the hard margin SVM solution does not make training errors, this convergence indicates that the soft margin SVM solutions may overfit if the regularization parameter is chosen too small.

Corollary 5.18 shows that a *linear* upper bound on A_2 of the form $A_2(\lambda) \leq c\lambda$ occurs if and only if $\mathcal{R}_{L,\mathrm{P}}$ has a global minimum in H. Now recall that Lemma 5.15 showed that such an upper bound on A_2 is optimal in the sense that every stronger upper bound of the form $A_2(\lambda) \leq \lambda h(\lambda)$, $\lambda \in [0,1]$, for some function $h : [0,1] \to [0,\infty)$ with $\lim_{\lambda\to 0^+} h(\lambda) = 0$, implies $A_2(\lambda) = 0$ for all $\lambda \geq 0$. Since the latter implies that 0 minimizes $\mathcal{R}_{L,\mathrm{P}}$ in H, we see that $f^*_{L,\mathrm{P},H}$ exists if and only if A_2 has an "optimal" upper bound.

Now assume that $f_{\mathrm{P},\lambda}$ exists for all $\lambda > 0$. By Lemma 5.15, we then find

$$\lambda \|f_{\mathrm{P},\lambda}\|_H^2 \leq A_2(\lambda) \leq \mathcal{R}_{L,\mathrm{P}}(0) - \mathcal{R}^*_{L,\mathrm{P},H}\,, \qquad \lambda > 0,$$

and hence we obtain $\lim_{\lambda\to\infty} f_{\mathrm{P},\lambda} = 0$. This justifies the notation $f_{\mathrm{P},\infty} := 0$. The next corollary shows the continuity of the (extended) function $\lambda \mapsto f_{\mathrm{P},\lambda}$.

Corollary 5.19. *Let $L : X \times Y \times \mathbb{R} \to [0,\infty)$ be a convex loss,* P *be a distribution on $X \times Y$ such that L is a* P*-integrable Nemitski loss, and H be the RKHS of a bounded measurable kernel on X. Then*

$$\lambda \mapsto f_{\mathrm{P},\lambda}$$

is a continuous map from $(0,\infty]$ to H, and $\lambda \mapsto \mathcal{R}_{L,\mathrm{P}}(f_{\mathrm{P},\lambda})$ is a continuous map from $(0,\infty]$ to $[0,\infty)$. Furthermore, if there exists an f^* minimizing $\mathcal{R}_{L,\mathrm{P}}$ in H, both maps are also defined and continuous at 0.

Proof. The first assertion is a direct consequence of Theorem 5.2 and Theorem 5.17, and the second assertion follows from combining the first assertion with Lemma 2.17. The last assertion follows from Corollary 5.18 and the notational convention $f_{\mathrm{P},0} := f^*_{L,\mathrm{P},H}$. □

Our next goal is to investigate whether the regularization term $\lambda\|f_{\mathrm{P},\lambda}\|_H^2$ and $A_2(\lambda)$ behave similarly for $\lambda \to 0$. To this end, we need some preparatory notions and results. Let us begin with the following definition, which introduces a differently regularized approximation error function.

Definition 5.20. *Let $L : X \times Y \times \mathbb{R} \to [0,\infty)$ be a convex loss, H be the RKHS of a measurable kernel on X, and* P *be a distribution on $X \times Y$ with $\mathcal{R}^*_{L,\mathrm{P},H} < \infty$. Then the* ***$\infty$-approximation error function*** *$A_\infty : [0,\infty) \to [0,\infty)$ is defined by*

$$A_\infty(\lambda) := \inf_{f\in\lambda^{-1}B_H} \mathcal{R}_{L,\mathrm{P}}(f) - \mathcal{R}^*_{L,\mathrm{P},H}\,, \qquad\qquad \lambda \geq 0.$$

Moreover, an element $f^{(\infty)}_{\mathrm{P},\lambda} \in \lambda^{-1}B_H$ *satisfying*

$$\mathcal{R}_{L,\mathrm{P}}(f^{(\infty)}_{\mathrm{P},\lambda}) = A_\infty(\lambda)$$

and $\|f^{(\infty)}_{\mathrm{P},\lambda}\| \leq \|f^*\|$ *for all* $f^* \in \lambda^{-1}B_H$ *with* $\mathcal{R}_{L,\mathrm{P}}(f^*) = A_\infty(\lambda)$ *is called a* ***minimal norm minimizer*** *of* $\mathcal{R}_{L,\mathrm{P}}$ *in* $\lambda^{-1}B_H$.

Note that, for RKHSs H satisfying $\mathcal{R}^*_{L,\mathrm{P},H} = \mathcal{R}^*_{L,\mathrm{P}}$, the value $A_\infty(\lambda)$ is usually called the **approximation error** of the set $\lambda^{-1}B_H$.

We will see later that there is an intimate relation between the functions A_2 and A_∞ and their minimizers. Before we go into details, we first present two results establishing the existence and uniqueness of minimal norm minimizers.

Lemma 5.21 (Uniqueness of minimal norm minimizers). *Let H be the RKHS of a measurable kernel on X, $L : X\times Y\times\mathbb{R} \to [0,\infty)$ be a convex loss, and* P *be a distribution on $X \times Y$ with $\mathcal{R}^*_{L,\mathrm{P},H} < \infty$. Then, for all $\lambda > 0$, there exists at most one minimal norm minimizer of $\mathcal{R}_{L,\mathrm{P}}$ in $\lambda^{-1}B_H$.*

Proof. Suppose that there are two different minimal norm minimizers $f_1^*, f_2^* \in \lambda^{-1}B_H$ of $\mathcal{R}_{L,\mathrm{P}}$ in $\lambda^{-1}B_H$. By the convexity of $\mathcal{R}_{L,\mathrm{P}}$, we then find that $\frac{1}{2}(f_1^* + f_2^*) \in \lambda^{-1}B_H$ also minimizes $\mathcal{R}_{L,\mathrm{P}}$. Moreover, we have (5.33), which contradicts our assumption that f^* is a minimal norm minimizer. □

The next lemma shows that the conditions ensuring the existence of general SVM solutions $f_{\mathrm{P},\lambda}$ also ensure the existence of minimal norm minimizers.

Lemma 5.22 (Existence of minimal norm minimizers). *Let* P *be a distribution on $X\times Y$, $L : X\times Y\times\mathbb{R} \to [0,\infty)$ be a convex* P*-integrable Nemitski loss, and H be the RKHS of a bounded measurable kernel on X. Then, for all $\lambda > 0$, there exists exactly one minimal norm minimizer $f^{(\infty)}_{\mathrm{P},\lambda} \in H$.*

Proof. We first show that there exists an $f^* \in \lambda^{-1}B_H$ that minimizes $\mathcal{R}_{L,\mathrm{P}}$ on $\lambda^{-1}B_H$. To this end, observe that $\mathcal{R}_{L,\mathrm{P}} : H \to [0,\infty)$ is continuous by Lemma 4.23 and Lemma 2.17. Therefore, the map $\mathcal{R} : H \to [0,\infty]$ defined by

$$\mathcal{R}(f) := \begin{cases} \mathcal{R}_{L,\mathrm{P}}(f) & \text{if } f \in \lambda^{-1}B_H \\ \infty & \text{otherwise} \end{cases}$$

is lower semi-continuous by Lemma A.2.8 and, in addition, $\mathcal{R}$ is also convex. Now recall that, since L is a P-integrable Nemitski loss, we have $\mathcal{R}_{L,\mathrm{P}}(0) < \infty$, and therefore every $f \in H$ with $\mathcal{R}(f) \leq \mathcal{R}_{L,\mathrm{P}}(0)$ must satisfy $\|f\| \leq \lambda^{-1}$. This shows that the set $\{f \in H : \mathcal{R}(f) \leq \mathcal{R}_{L,\mathrm{P}}(0)\}$ is non-empty and bounded, and consequently $\mathcal{R}$ has a global minimum at some $f^* \in \lambda^{-1}B_H$ by Theorem

A.6.9. Since $\mathcal{R}$ coincides with $\mathcal{R}_{L,\mathrm{P}}$ on $\lambda^{-1}B_H$, it is obvious that this f^* also minimizes $\mathcal{R}_{L,\mathrm{P}}$ on $\lambda^{-1}B_H$. In other words, the set

$$A := \left\{ f \in \lambda^{-1}B_H : \mathcal{R}_{L,\mathrm{P}}(f) \leq \mathcal{R}_{L,\mathrm{P}}(\tilde{f}) \text{ for all } \tilde{f} \in \lambda^{-1}B_H \right\}$$

is non-empty and, by the continuity of $\mathcal{R}_{L,\mathrm{P}} : H \to [0,\infty)$, we see that A is also closed. Moreover, $\|\cdot\|_H : H \to [0,\infty)$ is continuous. By applying Lemma A.2.8 and Theorem A.6.9 to the map $N : H \to [0,\infty]$ defined by

$$N(f) := \begin{cases} \|f\| & \text{if } f \in A \\ \infty & \text{otherwise,} \end{cases}$$

we hence see that $\|\cdot\|_H$ restricted to A has a minimum. □

Let us now compare the minimizers of A_2 and A_∞. We begin by showing that the existence of $f_{\mathrm{P},\lambda}$ implies that $f^{(\infty)}_{\mathrm{P},\gamma}$ exists for a suitable value γ.

Lemma 5.23. *Let $L : X \times Y \times \mathbb{R} \to [0,\infty)$ be a loss, H be the RKHS of a measurable kernel on X, and* P *be a distribution on $X \times Y$. Furthermore, assume that, for some $\lambda > 0$, there exists a unique $f_{\mathrm{P},\lambda}$. If $f_{\mathrm{P},\lambda} \neq 0$, then $f^{(\infty)}_{\mathrm{P},\gamma}$ exists for $\gamma := \|f_{\mathrm{P},\lambda}\|^{-1}$ and we have $f^{(\infty)}_{\mathrm{P},\gamma} = f_{\mathrm{P},\lambda}$.*

Proof. Let us first show that $f_{\mathrm{P},\lambda}$ minimizes $\mathcal{R}_{L,\mathrm{P}}$ on $\gamma^{-1}B_H$. To this end, assume the converse, i.e., that there exists an $f \in \gamma^{-1}B_H$ with

$$\mathcal{R}_{L,\mathrm{P}}(f) < \mathcal{R}_{L,\mathrm{P}}(f_{\mathrm{P},\lambda}) .$$

The definition of γ then gives $\|f\| \leq \gamma^{-1} = \|f_{\mathrm{P},\lambda}\|$, and hence we find

$$\lambda\|f\|^2 + \mathcal{R}_{L,\mathrm{P}}(f) < \lambda\|f_{\mathrm{P},\lambda}\|^2 + \mathcal{R}_{L,\mathrm{P}}(f_{\mathrm{P},\lambda}) . \tag{5.37}$$

Since the latter contradicts the definition of $f_{\mathrm{P},\lambda}$, we see that $f_{\mathrm{P},\lambda}$ minimizes $\mathcal{R}_{L,\mathrm{P}}$ on $\gamma^{-1}B_H$. Now assume that $f_{\mathrm{P},\lambda}$ is not a minimal norm minimizer. Then there exists an $f \in H$ with $\mathcal{R}_{L,\mathrm{P}}(f) = \mathcal{R}_{L,\mathrm{P}}(f_{\mathrm{P},\lambda})$ and $\|f\| < \|f_{\mathrm{P},\lambda}\|$. Since this leads again to the false (5.37), we obtain the assertion. □

If L is a convex, integrable Nemitski loss, we have already seen that the corresponding general SVM solutions exist and depend continuously on the regularization parameter. This leads to the following corollary.

Corollary 5.24. *Let $L : X \times Y \times \mathbb{R} \to [0,\infty)$ be a convex loss,* P *be a distribution on $X \times Y$ such that L is a* P*-integrable Nemitski loss, and H be the RKHS of a bounded measurable kernel on X with $\mathcal{R}^*_{L,\mathrm{P},H} < \mathcal{R}_{L,\mathrm{P}}(0)$. Then $\gamma : (0,\infty) \to (0,\infty)$ defined by $\gamma(\lambda) := \|f_{\mathrm{P},\lambda}\|^{-1}$, $\lambda > 0$, is a continuous map with*

$$f^{(\infty)}_{\mathrm{P},\gamma(\lambda)} = f_{\mathrm{P},\lambda} , \qquad \lambda > 0.$$

Consequently, $A_\infty : (0,\infty) \to (0,\infty)$ is an increasing and continuous map. Moreover, if there exists an f^* minimizing $\mathcal{R}_{L,\mathrm{P}}$ in H, we have $f^{(\infty)}_{\mathrm{P},\lambda} = f^*_{L,\mathrm{P},H}$ and $A_\infty(\lambda) = 0$ for all $0 \leq \lambda \leq \|f^*_{L,\mathrm{P},H}\|^{-1}$.

Proof. By Theorem 5.6 we know that $f_{\mathrm{P},\lambda} \neq 0$ for all $\lambda > 0$. The first assertion is hence a consequence of Corollary 5.19 and Lemma 5.23. The second assertion then follows from the first assertion and Lemma 2.17, and the third assertion is a consequence of the definition of A_∞. □

In order to appreciate the preceding corollary, let us assume that there does *not* exist an $f^* \in H$ minimizing $\mathcal{R}_{L,\mathrm{P}}$ in H. Then we know from Corollary 5.18 that $\lambda \mapsto \|f_{\mathrm{P},\lambda}\|$ is unbounded. Since $\lim_{\lambda\to\infty} f_{\mathrm{P},\lambda} = 0$, the intermediate value theorem then shows that for every $\gamma \in (0,\infty)$ there exists a $\lambda \in (0,\infty)$ with $\|f_{\mathrm{P},\lambda}\|^{-1} = \gamma$. Consequently, we obtain

$$\bigl\{f_{\mathrm{P},\lambda} : \lambda \in (0,\infty)\bigr\} = \bigl\{f^{(\infty)}_{\mathrm{P},\lambda} : \lambda \in (0,\infty)\bigr\}\,,$$

i.e., both approximation error functions produce the same set of minimizers, or **regularization path**, although they use a different form of regularization.

Let us finally compare the growth rates of the approximation error functions.

Theorem 5.25 (Quantitative comparison of A_2 and A_∞). *Let H be the RKHS of a measurable kernel on X, $L : X \times Y \times \mathbb{R} \to [0,\infty)$ be a loss, and P be a distribution on $X \times Y$ with $\mathcal{R}^*_{L,\mathrm{P},H} < \infty$. Then, for all $\lambda > 0$, the following statements are true:*

i) For every real number $\kappa \geq A_\infty(\lambda)$, we have

$$A_2\bigl(\lambda^2\kappa\bigr) \leq 2\kappa\,. \tag{5.38}$$

ii) For every real number $\gamma > A_2(\lambda)$, we have

$$A_\infty\bigl(\lambda^{1/2}\gamma^{-1/2}\bigr) \;\leq\; \gamma\,. \tag{5.39}$$

In addition, if $f_{\mathrm{P},\lambda}$ exists, then (5.39) also holds for $\gamma := A_2(\lambda)$.

Proof. i). For $\varepsilon > 0$ there exists an $f_\varepsilon \in \lambda^{-1}B_H$ with $\mathcal{R}_{L,\mathrm{P}}(f_\varepsilon) - \mathcal{R}^*_{L,\mathrm{P},H} \leq \kappa + \varepsilon$. If $f_\varepsilon \neq 0$, we have $\lambda \leq \|f_\varepsilon\|^{-1}$, and hence we obtain $\lambda^2\kappa \leq \|f_\varepsilon\|^{-2}\kappa =: \lambda_\varepsilon$. By the monotonicity of $A_2(\,\cdot\,)$, the latter yields

$$A_2\bigl(\lambda^2\kappa\bigr) \;\leq\; A_2(\lambda_\varepsilon) \;\leq\; \lambda_\varepsilon\|f_\varepsilon\|^2 + \mathcal{R}_{L,\mathrm{P}}(f_\varepsilon) - \mathcal{R}^*_{L,\mathrm{P},H} \;\leq\; 2\kappa + \varepsilon\,.$$

Letting $\varepsilon \to 0$, we then obtain the assertion. The case $f_\varepsilon = 0$ can be shown analogously by setting $\lambda_\varepsilon := \lambda^2\kappa$.

ii). Since $\gamma > A_2(\lambda)$, there exists an $f \in H$ with

$$\lambda\|f\|^2 + \mathcal{R}_{L,\mathrm{P}}(f) - \mathcal{R}^*_{L,\mathrm{P},H} \leq \gamma,$$

and since $\mathcal{R}_{L,\mathrm{P}}(f) - \mathcal{R}^*_{L,\mathrm{P},H} \geq 0$, this f satisfies $\lambda\|f\|^2 \leq \gamma$. The latter gives $\|f\| \leq (\gamma/\lambda)^{1/2} =: \kappa^{-1}$, and hence we find

$$A_\infty(\kappa) \;\leq\; \mathcal{R}_{L,\mathrm{P}}(f) - \mathcal{R}^*_{L,\mathrm{P},H} \;\leq\; \lambda\|f\|^2 + \mathcal{R}_{L,\mathrm{P}}(f) - \mathcal{R}^*_{L,\mathrm{P},H} \;\leq\; \gamma\,.$$

Finally, if $f_{\mathrm{P},\lambda}$ exists, we have $\lambda\|f_{\mathrm{P},\lambda}\|^2 + \mathcal{R}_{L,\mathrm{P}}(f_{\mathrm{P},\lambda}) - \mathcal{R}^*_{L,\mathrm{P},H} = A_2(\lambda)$, and hence repeating the reasoning above gives the second assertion. □

If $\lambda \mapsto A_\infty(\lambda)$ behaves polynomially, then Theorem 5.25 can be used to show that the asymptotic behavior of $\lambda \mapsto A_2(\lambda)$ is polynomial and completely determined by that of $\lambda \mapsto A_\infty(\lambda)$. We refer to Exercise 5.11 for details. For non-polynomial behavior, however, this is no longer true. Indeed, if there exists a function minimizing $\mathcal{R}_{L,\mathrm{P}}$ in H, then Lemma 5.15 and Corollary 5.18 showed that

$$\lambda A_2(1) \leq A_2(\lambda) \leq \lambda \|f^*_{L,\mathrm{P},H}\|^2_H\,, \qquad \lambda \in [0,1],$$

whereas Corollary 5.24 showed that $A_\infty(\lambda) = 0$ for all $0 \leq \lambda \leq \|f^*_{L,\mathrm{P},H}\|^{-1}_H$.

Let us finally discuss how the behavior of $A_2(\lambda)$ for $\lambda \to 0$ influences the behavior of $\|f_{\mathrm{P},\lambda}\|$ for $\lambda \to 0$. To this end, let us recall that we found

$$\|f_{\mathrm{P},\lambda}\|_H \leq \sqrt{\frac{\mathcal{R}_{L,\mathrm{P}}(0)}{\lambda}} \tag{5.40}$$

at the beginning of Section 5.1. Unfortunately, this bound is *always* suboptimal for $\lambda \to 0$. Indeed, we have $\lambda\|f_{\mathrm{P},\lambda}\|^2 \leq A_2(\lambda)$, and hence we obtain

$$\|f_{\mathrm{P},\lambda}\| \leq \sqrt{\frac{A_2(\lambda)}{\lambda}}\,, \qquad \lambda > 0. \tag{5.41}$$

Since $A_2(\lambda) \to 0$ for $\lambda \to 0$, the latter bound is strictly sharper than (5.40). This observation will be of great importance when investigating the stochastic properties of empirical SVM solutions in Chapter 7.

5.5 Approximation Error of RKHSs

We saw in (5.32) that, for vanishing regularization parameter, the risk of the general SVM solution tends to the minimal risk $\mathcal{R}^*_{L,\mathrm{P},H}$ of the used RKHS H. However, our ultimate interest is to investigate whether the risk of the empirical SVM tends to the Bayes risk $\mathcal{R}^*_{L,\mathrm{P}}$. Following the intuition that the risk of the empirical SVM solution $f_{\mathrm{D},\lambda}$ is close to that of the corresponding infinite-sample SVM solution $f_{\mathrm{P},\lambda}$, we can conjecture that $\mathcal{R}_{L,\mathrm{P}}(f_{\mathrm{D},\lambda})$ is close to $\mathcal{R}^*_{L,\mathrm{P},H}$. Moreover, in Section 6.4 we will show that this conjecture is indeed correct, and consequently we need to know that

$$\mathcal{R}^*_{L,\mathrm{P},H} = \mathcal{R}^*_{L,\mathrm{P}}$$

in order to show that $\mathcal{R}_{L,\mathrm{P}}(f_{\mathrm{D},\lambda})$ is close to $\mathcal{R}^*_{L,\mathrm{P}}$. The goal of this section is to establish necessary and sufficient conditions on H for this equality to hold.

Let us begin with a rather technical observation. In Chapter 3, we often used *complete* measurable spaces X in order to ensure the measurability of many considered functions such as Bayes decision functions. On the other hand, we considered continuous kernels with "large" RKHSs over (compact) metric spaces X in Section 4.6. However, in general, the Borel σ-algebra $\mathcal{B}(X)$

is *not* complete, and hence one may ask which σ-algebra we should use for computing the Bayes risk. Fortunately, it turns out by (A.8) that we have

$$\mathcal{R}^*_{L,\mathrm{P}} = \mathcal{R}^*_{L,\hat{\mathrm{P}}}\,,$$

where $\hat{\mathrm{P}}$ is the extension of P to the P-completion $\mathcal{B}_{\mathrm{P}}(X)$ of $\mathcal{B}(X)$. In other words, the Bayes risk does not change when using these different σ-algebras, and hence we can always choose the σ-algebra that best fits our needs.

In the following, we will often use intermediate approximation results where the space H in $\mathcal{R}^*_{L,\mathrm{P},H}$ is replaced by another set of measurable functions. Let us formalize this idea by the following definition.

Definition 5.26. *Let $L : X \times Y \times \mathbb{R} \to [0,\infty)$ be a loss, P be a distribution on $X \times Y$, and $F \subset \mathcal{L}_0(X)$ be a set of measurable functions. Then the **minimal L-risk** over F is*

$$\mathcal{R}^*_{L,\mathrm{P},F} := \inf\bigl\{\mathcal{R}_{L,\mathrm{P}}(f) : f \in F\bigr\}\,.$$

Now our first result shows that, for integrable Nemitski losses, the Bayes risk can be approximated by *bounded* measurable functions.

Proposition 5.27. *Let $L : X \times Y \times \mathbb{R} \to [0,\infty)$ be a loss and P be a distribution on $X \times Y$ such that L is a P-integrable Nemitski loss. Then we have*

$$\mathcal{R}^*_{L,\mathrm{P},L_\infty(\mathrm{P}_X)} = \mathcal{R}^*_{L,\mathrm{P}}\,.$$

Proof. Let us fix a measurable $f : X \to \mathbb{R}$ with $\mathcal{R}_{L,\mathrm{P}}(f) < \infty$. Then the functions $f_n := \mathbf{1}_{\{|f|\le n\}} f$, $n \ge 1$, are bounded, and an easy calculation yields

$$\begin{aligned}\bigl|\mathcal{R}_{L,\mathrm{P}}(f_n) - \mathcal{R}_{L,\mathrm{P}}(f)\bigr| &\le \int_{\{|f|>n\}\times Y} \bigl|L(x,y,0) - L(x,y,f(x))\bigr|\,d\mathrm{P}(x,y)\\ &\le \int_{\{|f|>n\}\times Y} b(x,y) + h(0) + L\bigl(x,y,f(x)\bigr)\,d\mathrm{P}(x,y)\end{aligned}$$

for all $n \ge 1$. In addition, the integrand in the last integral is integrable since $\mathcal{R}_{L,\mathrm{P}}(f) < \infty$ and $b \in \mathcal{L}_1(\mathrm{P})$, and consequently Lebesgue's theorem yields $\mathcal{R}_{L,\mathrm{P}}(f_n) \to \mathcal{R}_{L,\mathrm{P}}(f)$ for $n \to \infty$. From this, we easily get the assertion. □

Using Proposition 5.27, we can now establish our first condition on F, which ensures $\mathcal{R}^*_{L,\mathrm{P},F} = \mathcal{R}^*_{L,\mathrm{P}}$.

Theorem 5.28 (Approximation by pointwise dense sets). *Let P be a distribution on $X \times Y$, $L : X \times Y \times \mathbb{R} \to [0,\infty)$ be a continuous P-integrable Nemitski loss, and $F \subset L_\infty(\mathrm{P}_X)$. Assume that for all $g \in L_\infty(\mathrm{P}_X)$ there exists a sequence $(f_n) \subset F$ such that $\sup_{n\ge1}\|f_n\|_\infty < \infty$ and*

$$\lim_{n\to\infty} f_n(x) \;=\; g(x)$$

for P_X-almost all $x \in X$. Then we have $\mathcal{R}^*_{L,\mathrm{P},F} = \mathcal{R}^*_{L,\mathrm{P}}$.

Proof. By Proposition 5.27, we know that $\mathcal{R}^*_{L,\mathrm{P},L_\infty(\mathrm{P}_X)} = \mathcal{R}^*_{L,\mathrm{P}}$, and since $F \subset L_\infty(\mathrm{P}_X)$ we also have $\mathcal{R}^*_{L,\mathrm{P},F} \geq \mathcal{R}^*_{L,\mathrm{P},L_\infty(\mathrm{P}_X)}$. In order to show the converse inequality, we fix a $g \in L_\infty(\mathrm{P}_X)$. Let $(f_n) \subset H$ be a sequence of functions according to the assumptions of the theorem. Lemma 2.17 then yields $\mathcal{R}_{L,\mathrm{P}}(f_n) \to \mathcal{R}_{L,\mathrm{P}}(g)$, and hence we find $\mathcal{R}^*_{L,\mathrm{P},F} \leq \mathcal{R}^*_{L,\mathrm{P},L_\infty(\mathrm{P}_X)}$. □

With the help of Theorem 5.28, we now show that the RKHSs of universal kernels approximate the Bayes risk of continuous, integrable Nemitski losses.

Corollary 5.29. *Let X be a compact metric space, H be the RKHS of a universal kernel on X, P be a distribution on $X \times Y$, and $L : X \times Y \times \mathbb{R} \to [0,\infty)$ be a continuous* P*-integrable Nemitski loss. Then we have*

$$\mathcal{R}^*_{L,\mathrm{P},H} = \mathcal{R}^*_{L,\mathrm{P}}\,.$$

Proof. Let us fix a $g \in L_\infty(\mathrm{P}_X)$. By Theorem A.5.25, there then exists a sequence $(g_n) \subset C(X)$ with $\|g_n - g\|_1 \to 0$ for $n \to \infty$. By clipping g_n at $\pm\|g\|_\infty$ and considering a suitable subsequence, we see that we can assume without loss of generality that $\|g_n\|_\infty \leq \|g\|_\infty$ for all $n \geq 1$ and $g_n(x) \to g(x)$ for P_X-almost all $x \in X$. Moreover, the universality of H gives a sequence $(f_n) \subset H$ with $\|f_n - g_n\|_\infty \leq 1/n$ for all $n \geq 1$. Since this yields both $\|f_n\|_\infty \leq 1 + \|g\|_\infty$ for all $n \geq 1$ and $f_n(x) \to g(x)$ for P_X-almost all $x \in X$, we obtain the assertion by Theorem 5.28. □

The next theorem, which is a consequence of Theorem 4.61, provides a result similar to Corollary 5.29 for discrete input spaces.

Theorem 5.30. *Let X be a countable set and k be a bounded kernel on X that satisfies $k(\,\cdot\,,x) \in c_0(X)$ for all $x \in X$ and*

$$\sum_{x,x' \in X} k(x,x')f(x)f(x') > 0 \tag{5.42}$$

for all $f \in \ell_1(X)$ with $f \neq 0$. Then the RKHS H of k satisfies

$$\mathcal{R}^*_{L,\mathrm{P},H} = \mathcal{R}^*_{L,\mathrm{P}}$$

for all closed $Y \subset \mathbb{R}$, all distributions P *on $X \times Y$, and all continuous,* P*-integrable Nemitski losses $L : X \times Y \times \mathbb{R} \to [0,\infty)$.*

Proof. We have already seen in Theorem 4.61 that H is dense in $c_0(X)$. By Lemma 2.17, we hence find $\mathcal{R}^*_{L,\mathrm{P},H} = \mathcal{R}^*_{L,\mathrm{P},c_0(X)}$. Therefore it remains to show

$$\mathcal{R}^*_{L,\mathrm{P},c_0(X)} = \mathcal{R}^*_{L,\mathrm{P}}\,. \tag{5.43}$$

To this end, let ν be the counting measure on X and $h : X \to [0,1]$ be the map that satisfies $\mathrm{P}_X = h\nu$. In addition, recall that we have $\mathcal{R}_{L,\mathrm{P}}(0) < \infty$

since L is a P-integrable Nemitski loss. Given an $\varepsilon > 0$, there hence exists a *finite* set $A \subset X$ with

$$\sum_{x \in X \setminus A} h(x) \int_Y L(x, y, 0) d\mathrm{P}(y|x) \leq \varepsilon\,.$$

In addition, there exists a $g : X \to \mathbb{R}$ with $\mathcal{R}_{L,\mathrm{P}}(g) \leq \mathcal{R}^*_{L,\mathrm{P}} + \varepsilon$. Let us define $f := \mathbf{1}_A g$. Then we have $f \in c_0(X)$ and

$$\begin{aligned}\mathcal{R}_{L,\mathrm{P}}(f) &= \sum_{x \in A} h(x) \int_Y L\big(x, y, g(x)\big) d\mathrm{P}(y|x) + \sum_{x \in X \setminus A} h(x) \int_Y L(x, y, 0) d\mathrm{P}(y|x) \\ &\leq \mathcal{R}_{L,\mathrm{P}}(g) + \varepsilon\,.\end{aligned}$$

From this we easily infer (5.43). □

If L is actually a continuous, P-integrable Nemitski loss of some order $p \in [1, \infty)$, we have already seen in Lemma 2.17 that the risk functional $\mathcal{R}_{L,\mathrm{P}}$ is even continuous on $L_p(\mathrm{P}_X)$. This leads to the following result.

Theorem 5.31 (Approximation by p-integrable functions). *Let* P *be a distribution on* $X \times Y$ *and* $L : X \times Y \times \mathbb{R} \to [0, \infty)$ *be a continuous* P*-integrable Nemitski loss of order* $p \in [1, \infty)$*. Then, for every dense* $F \subset L_p(\mathrm{P}_X)$*, we have*

$$\mathcal{R}^*_{L,\mathrm{P},F} = \mathcal{R}^*_{L,\mathrm{P}}\,.$$

Proof. Since $L_\infty(\mathrm{P}_X) \subset L_p(\mathrm{P}_X)$, we have $\mathcal{R}^*_{L,\mathrm{P},L_p(\mathrm{P}_X)} = \mathcal{R}^*_{L,\mathrm{P}}$ by Proposition 5.27. Now the assertion easily follows from the denseness of F in $L_p(\mathrm{P}_X)$ and the continuity of $\mathcal{R}_{L,\mathrm{P}} : L_p(\mathrm{P}_X) \to [0, \infty)$ established in Lemma 2.17. □

Recall that we have shown in Theorem 4.63 that the RKHSs H_γ of the Gaussian RBF kernels are dense in $L_p(\mu)$ for all $p \in [1, \infty)$ and all distributions μ on $\mathbb{R}^d$. With the help of the preceding theorem, it is then easy to establish $\mathcal{R}^*_{L,\mathrm{P},H_\gamma} = \mathcal{R}^*_{L,\mathrm{P}}$ for almost all of the margin-based and distance-based loss functions considered in Sections 2.3 and 2.4. The details are left to the reader as an additional exercise.

Let us now derive some *necessary* conditions for $\mathcal{R}^*_{L,\mathrm{P},F} = \mathcal{R}^*_{L,\mathrm{P}}$. To this end, we need the following lemma.

Lemma 5.32. *Let* X *be a measurable space and* μ *be a probability measure on* X*. Assume that we have a subspace* $F \subset L_\infty(\mu)$ *such that for all measurable* $A \subset X$ *there exists a sequence* $(f_n) \subset F$ *with*

$$\lim_{n \to \infty} f_n(x) = \mathbf{1}_A(x)$$

for μ*-almost all* $x \in X$*. Then, for all* $g \in L_\infty(\mu)$*, there exists a sequence* $(f_n) \subset F$ *with* $\lim_{n\to\infty} f_n(x) = g(x)$ *for* μ*-almost all* $x \in X$*.*

Proof. Observe that, for measurable step functions $g \in L_\infty(\mu)$, the assertion immediately follows from the fact that F is a vector space. Let us now fix an arbitrary $g \in L_\infty(\mu)$. For $n \geq 1$, there then exists a measurable step function $g_n \in L_\infty(\mu)$ with $\|g_n - g\|_\infty \leq 1/n$. For this g_n, there then exists a sequence $(f_{m,n})_{m\geq 1} \subset F$ with $\lim_{m\to\infty} f_{m,n}(x) = g_n(x)$ for μ-almost all $x \in X$. By Egorov's Theorem A.3.8, we find a measurable $A_n \subset X$ with $\mu(X\backslash A_n) \leq 1/n$ and

$$\lim_{m\to\infty} \|(f_{m,n} - g_n)_{|A_n}\|_\infty = 0 \,.$$

Consequently, there is an index $m_n \geq 1$ with $\|(f_{m_n,n} - g_n)_{|A_n}\|_\infty \leq 1/n$. By putting all estimates together, we now obtain

$$\mu\biggl(\Bigl\{x \in X : \bigl|f_{m_n,n}(x) - g(x)\bigr| \leq \frac{2}{n}\Bigr\}\biggr) \geq 1 - \frac{1}{n}\,, \qquad\qquad n \geq 1.$$

Therefore the sequence $(f_{m_n,n})_{n\geq 1}$ converges to g in probability μ, and hence there exists a subsequence of it that converges to g almost surely. □

With the help of the preceding lemma, we can now formulate a necessary condition for *convex* supervised loss functions. For its formulation, we need to recall from Section 3.1 that $\mathcal{M}_{L,\mathrm{Q}}(0^+)$ denotes the set of exact minimizers of the inner L-risk of Q. Moreover, recall Definition 3.5, where we said that a distribution P on $X \times Y$ is of type $\mathcal{Q}$ if $\mathcal{Q}$ is a set of distributions on Y and $\mathrm{P}(\,\cdot\,|x) \in \mathcal{Q}$ for P_X-almost all $x \in X$.

Theorem 5.33 (Pointwise denseness is necessary). *Let $L : Y \times \mathbb{R} \to [0,\infty)$ be a convex supervised loss for which there exist two distributions Q_1 and Q_2 on Y and $t_1^*, t_2^* \in \mathbb{R}$ such that $t_1^* \neq t_2^*$, $\mathcal{M}_{L,\mathrm{Q}_1}(0^+) = \{t_1^*\}$, and $\mathcal{M}_{L,\mathrm{Q}_2}(0^+) = \{t_2^*\}$. Furthermore, let X be a measurable space and μ be a distribution on X. Assume that $F \subset L_\infty(\mu)$ is a subspace with*

$$\mathcal{R}^*_{L,\mathrm{P},F} = \mathcal{R}^*_{L,\mathrm{P}} \tag{5.44}$$

for all $\{\mathrm{Q}_1, \mathrm{Q}_2\}$-type distributions P on $X \times Y$ with $\mathrm{P}_X = \mu$. Then, for all $g \in L_\infty(\mu)$, there exists a sequence $(f_n) \subset F$ such that $\lim_{n\to\infty} f_n(x) = g(x)$ for μ-almost all $x \in X$.

Proof. Let $\mathcal{A}$ be the σ-algebra of the measurable space X. We fix an $A_1 \in \mathcal{A}$ and write $A_2 := \emptyset$. Let us define two distributions P_1 and P_2 on $X \times Y$ by

$$\mathrm{P}_i(\,\cdot\,|x) := \begin{cases} \mathrm{Q}_1 & \text{if } x \in A_i \\ \mathrm{Q}_2 & \text{if } x \in X\backslash A_i \end{cases}$$

and $(\mathrm{P}_i)_X := \mu$ for $i = 1, 2$. Our assumptions on Q_1 and Q_2 guarantee $\mathcal{C}^*_{L,\mathrm{Q}_1} < \infty$ and $\mathcal{C}^*_{L,\mathrm{Q}_2} < \infty$ by Lemma 3.10, and hence we find $\mathcal{R}^*_{L,\mathrm{P}_i} < \infty$ for $i = 1, 2$. Moreover, every Bayes decision function of $\mathcal{R}_{L,\mathrm{P}_i}$, $i = 1, 2$, has μ-almost surely the form

$$f^*_{L,\mathrm{P}_i} := t_1^* \mathbf{1}_{A_i} + t_2^* \mathbf{1}_{X\setminus A_i}\,, \qquad\qquad i=1,2.$$

Now, (5.44) yields $\mathcal{R}^*_{L,\mathrm{P}_i,F} = \mathcal{R}^*_{L,\mathrm{P}_i}$, $i=1,2$, and hence there are sequences $(f_n^{(1)}) \subset F$ and $(f_n^{(2)}) \subset F$ with

$$\lim_{n\to\infty} \mathcal{R}_{L,\mathrm{P}_i}(f_n^{(i)}) = \mathcal{R}^*_{L,\mathrm{P}_i}\,, \qquad\qquad i=1,2. \tag{5.45}$$

Recalling that convex supervised losses are self-calibrated, Corollary 3.62 then shows for $i=1,2$ that

$$\lim_{n\to\infty} f_n^{(i)} = f^*_{L,\mathrm{P}_i} \tag{5.46}$$

in probability $\hat\mu$, where $\hat\mu$ is the extension of μ to the μ-completed σ-algebra $\mathcal{A}_\mu$, see Lemma A.3.3. Now observe that all functions in (5.46) are $\mathcal{A}$-measurable, and hence (5.46) actually holds in probability μ. Consequently, there exist subsequences $(f_{n_j}^{(1)})$ and $(f_{n_j}^{(2)})$ for which (5.46) holds μ-almost surely. For

$$f_j := \frac{1}{t_1^* - t_2^*}\bigl(f_{n_j}^{(1)} - f_{n_j}^{(2)}\bigr)\,, \qquad\qquad j\geq 1,$$

we then have $f_j \in F$, and in addition our construction yields

$$\lim_{j\to\infty} f_j = \frac{1}{t_1^*-t_2^*}\bigl(f^*_{L,\mathrm{P}_1} - f^*_{L,\mathrm{P}_2}\bigr) = \frac{1}{t_1^*-t_2^*}\bigl(t_1^*\mathbf{1}_{A_1} + t_2^*\mathbf{1}_{X\setminus A_1} - t_2^*\mathbf{1}_X\bigr) = \mathbf{1}_{A_1}$$

μ-almost surely. By Lemma 5.32, we thus obtain the assertion. □

For RKHSs, Theorem 5.33 has an interesting implication, which is presented in the following corollary.

Corollary 5.34 (Kernels should be strictly positive definite). *Let L be a loss function satisfying the assumptions of Theorem 5.33. Furthermore, let X be a measurable space and k be a measurable kernel on X whose RKHS H satisfies $\mathcal{R}^*_{L,\mathrm{P},H} = \mathcal{R}^*_{L,\mathrm{P}}$ for all $\{\mathrm{Q}_1,\mathrm{Q}_2\}$-type distributions P on $X\times Y$. Then k is strictly positive definite.*

Proof. Let $x_1,\dots,x_n \in X$ be mutually different points and μ be the associated empirical distribution. Obviously, it suffices to show that the kernel matrix $K := (k(x_i,x_j))$ has full rank. Let us assume the converse, i.e., we assume that there exists an $y\in\mathbb{R}^n$ with $K\alpha \neq y$ for all $\alpha\in\mathbb{R}^n$. Since $K\mathbb{R}^n$ is closed, there then exists an $\varepsilon>0$ with $\|K\alpha - y\|_\infty \geq \varepsilon$ for all $\alpha\in\mathbb{R}^n$. Now note that every $\bar f\in H$ that is orthogonal to $\operatorname{span}\{k(\,\cdot\,,x_i) : i=1,\dots,n\}$ satisfies

$$\bar f(x_j) = \langle \bar f, k(\,\cdot\,,x_j)\rangle_H = 0\,, \qquad\qquad j=1,\dots,n.$$

By decomposing H into $\operatorname{span}\{k(\,\cdot\,,x_i) : i=1,\dots,n\}$ and its orthogonal complement, we consequently see that for every $f\in H$ there is an $\alpha\in\mathbb{R}^n$ with

$$f(x_j) = \sum_{i=1}^n \alpha_i k(x_j,x_i)\,, \qquad\qquad j=1,\dots,n,$$

and hence for all $f \in H$ there is an index $j \in \{1, \dots, n\}$ with $|f(x_j) - y_j| > \varepsilon$. On the other hand, k is bounded on $\{1, \dots, n\}$, and hence Theorem 5.33 gives a sequence $(f_n) \subset H$ with $f_n(x_i) \to y_i$ for all $i \in \{1, \dots, n\}$. From this we easily find a contradiction. □

In order to illustrate the previous results, let us assume for a moment that we have a distribution μ on X and a convex loss $L : Y \times \mathbb{R} \to [0, \infty)$ that satisfy the assumptions of both Theorem 5.28 and Theorem 5.33. In addition, let $F \subset L_\infty(\mu)$ be a subspace. Now assume that we are interested in the question of whether

$$\mathcal{R}^*_{L,\mathrm{P},F} \;=\; \mathcal{R}^*_{L,\mathrm{P}} \tag{5.47}$$

holds for a reasonably large class of distributions P with $\mathrm{P}_X = \mu$. Theorem 5.33 then shows that a *necessary* condition for (5.47) to hold is that F be "dense" in $L_\infty(\mu)$ with respect to the μ-almost sure convergence[2], i.e., every $g \in L_\infty(\mu)$ is the μ-almost sure limit of a suitable sequence $(f_n) \subset F$. However, the *sufficient* condition of Theorem 5.28 for (5.47) to hold requires that there actually be such a sequence (f_n) that is *uniformly* bounded. In other words, there is a gap between the necessary and the sufficient conditions. Now recall that in Theorem 5.31 we have already weakened the sufficient condition for a more restricted class of loss functions. In the following, we will present a stronger necessary condition by making additional assumptions on the distributions Q_1 and Q_2 of Theorem 5.33, so in the end we obtain characterizations on F for (5.47) to hold for a variety of important loss functions. Let us begin with the following simple lemma.

Lemma 5.35. *Let X be a measurable space, μ be a probability measure on X, and $q > 0$. Assume that we have a subspace $F \subset L_q(\mu)$ such that for all measurable $A \subset X$ there exists a sequence $(f_n) \subset F$ with*

$$\lim_{n\to\infty} \|f_n - \mathbf{1}_A\|_{L_q(\mu)} = 0\,. \tag{5.48}$$

Then F is dense in $L_q(\mu)$.

Proof. If $g \in L_q(\mu)$ is a measurable step function, there obviously exists a sequence $(f_n) \subset F$ with $\lim_{n\to\infty} \|f_n - g\|_q = 0$. Moreover, if $g \in L_q(\mu)$ is bounded and n is an integer, there exists a measurable step function g_n with $\|g_n - g\|_\infty \leq 1/n$. In addition, we have just seen that there exists an $f_n \in F$ with $\|f_n - g_n\|_q \leq 1/n$, and hence we find $\lim_{n\to\infty} \|f_n - g\|_q = 0$. Finally, for general $g \in L_q(\mu)$, we find an approximating sequence by first approximating g with the bounded measurable functions $g_n := \mathbf{1}_{|g|\leq n} g$, $n \geq 1$, and then approximating these g_n with suitable functions $f_n \in F$. □

[2] Note that in general there is no topology generating the almost sure convergence, and thus "dense" is not really defined. However, the almost sure convergence in Theorem 5.33 and Theorem 5.28 can be replaced by the convergence in probability (see Exercise 5.12), and since this convergence originates from a metric, we then have a precise meaning of "dense".

With the help of the preceding lemma, we can now establish a stronger necessary condition on F for $\mathcal{R}^*_{L,\mathrm{P},F} = \mathcal{R}^*_{L,\mathrm{P}}$ to hold. For its formulation, we need to recall the self-calibration function defined in (3.67).

Theorem 5.36 (Denseness in $L_q(\mu)$ is necessary). *Let $L : Y \times \mathbb{R} \to [0, \infty)$ be a convex supervised loss such that there exist two distributions Q_1 and Q_2 on Y and t_1^*, $t_2^* \in \mathbb{R}$ with $t_1^* \neq t_2^*$, $\mathcal{M}_{L,\mathrm{Q}_1}(0^+) = \{t_1^*\}$ and $\mathcal{M}_{L,\mathrm{Q}_2}(0^+) = \{t_2^*\}$. In addition, assume that their self-calibration functions satisfy*

$$\delta_{\max,\breve{L},L}(\varepsilon, \mathrm{Q}_i) \geq B\varepsilon^q\,, \qquad \varepsilon > 0,\ i = 1, 2,$$

for some constants $B > 0$ and $q > 0$. Furthermore, let X be a measurable space, μ be a distribution on X, and $F \subset L_q(\mu)$ be a subspace with

$$\mathcal{R}^*_{L,\mathrm{P},F} = \mathcal{R}^*_{L,\mathrm{P}}$$

for all $\{\mathrm{Q}_1, \mathrm{Q}_2\}$-type distributions P on $X \times Y$ with $\mathrm{P}_X = \mu$. Then F is dense in $L_q(\mu)$.

Proof. Following the argument used in the proof of Theorem 5.33, we may assume without loss of generality that X is a complete measurable space. Let us now fix a measurable $A_1 \subset X$ and write $A_2 := \emptyset$. Furthermore, we define the distributions P_i, their Bayes decision functions f^*_{L,P_i}, and the approximating sequences $(f_n^{(i)}) \subset F$, $i = 1, 2$, as in the proof of Theorem 5.33. Then (5.45) together with (3.69) for $p := \infty$ yields

$$\lim_{n\to\infty} \|f_n^{(i)} - f^*_{L,\mathrm{P}_i}\|_{L_q(\mu)} = 0\,.$$

For $f_n := \frac{1}{t_1^* - t_2^*}\bigl(f_n^{(1)} - f_n^{(2)}\bigr), n \geq 1$, we then obtain $\lim_{n\to\infty} \|f_n - \mathbf{1}_{A_1}\|_{L_q(\mu)} = 0$, and hence we obtain the assertion by Lemma 5.35. □

By combining Theorem 5.31 with Theorem 5.36, we now obtain the following characterization of subspaces F satisfying $\mathcal{R}^*_{L,\mathrm{P},F} = \mathcal{R}^*_{L,\mathrm{P}}$.

Corollary 5.37 (Characterization). *Let $L : Y \times \mathbb{R} \to [0, \infty)$ be a convex supervised Nemitski loss of order $p \in [1, \infty)$, i.e.*

$$L(y, t) \leq b(y) + c|t|^p\,, \qquad (y, t) \in Y \times \mathbb{R}, \tag{5.49}$$

for a suitable constant $c > 0$ and a measurable function $b : Y \to [0, \infty)$. Furthermore, let Q_1 and Q_2 be distributions on Y with $b \in L_1(\mathrm{Q}_1) \cap L_1(\mathrm{Q}_2)$ and t_1^*, $t_2^* \in \mathbb{R}$ with $t_1^* \neq t_2^*$, $\mathcal{M}_{L,\mathrm{Q}_1}(0^+) = \{t_1^*\}$ and $\mathcal{M}_{L,\mathrm{Q}_2}(0^+) = \{t_2^*\}$. In addition, assume that their self-calibration functions satisfy

$$\delta_{\max,\breve{L},L}(\varepsilon, \mathrm{Q}_i) \geq B\,\varepsilon^p\,, \qquad \varepsilon > 0,\ i = 1, 2,$$

for some constant $B > 0$. Furthermore, let X be a measurable space, μ be a distribution on X, and $F \subset L_p(\mu)$ be a subspace. Then the following statements are equivalent:

i) F is dense in $L_p(\mu)$.

ii) For all distributions P on $X \times Y$ with $\mathrm{P}_X = \mu$ for which L is a P-integrable Nemitski loss of order $p \in [1, \infty)$, we have $\mathcal{R}^*_{L,\mathrm{P},F} = \mathcal{R}^*_{L,\mathrm{P}}$.

iii) For all $\{\mathrm{Q}_1, \mathrm{Q}_2\}$-type distributions P on $X \times Y$ with $\mathrm{P}_X = \mu$, we have $\mathcal{R}^*_{L,\mathrm{P},F} = \mathcal{R}^*_{L,\mathrm{P}}$.

Proof. *i)* $\Rightarrow$ *ii)*. Use Theorem 5.31.

ii) $\Rightarrow$ *iii)*. By (5.49) and $b \in L_1(\mathrm{Q}_1) \cap L_1(\mathrm{Q}_2)$, we see that L is a P-integrable Nemitski loss of order $p \in [1, \infty)$ for all $\{\mathrm{Q}_1, \mathrm{Q}_2\}$-type distributions P on $X \times Y$.

iii) $\Rightarrow$ *i)*. Use Theorem 5.36. □

Let us now illustrate that many important losses satisfy the assumptions of Corollary 5.37. For some more examples, we refer to Exercise 5.13.

Example 5.38. For $p \geq 1$, let L be the p-th **power absolute distance loss** defined in Example 2.39. Moreover, let $\mathrm{Q}_1 := \delta_{\{y_1\}}$ and $\mathrm{Q}_1 := \delta_{\{y_2\}}$ be two Dirac distributions on $\mathbb{R}$ with $y_1 \neq y_2$. Then L, Q_1, and Q_2 satisfy the assumptions of Corollary 5.37.

In order to see this, recall that L is a Nemitski loss of order p by Example 2.39 and Lemma 2.36. Furthermore, for the Dirac measure $\mathrm{Q} := \delta_{\{y\}}$ at an arbitrary $y \in \mathbb{R}$, we have $\mathcal{C}_{L,\mathrm{Q}}(t) = |t - y|^p$, $t \in \mathbb{R}$. Consequently, we have $\mathcal{C}^*_{L,\mathrm{Q}} = 0$ and $\mathcal{M}_{L,\mathrm{Q}}(0^+) = \{y\}$. With these equalities, it is easy to check that the self-calibration function of L is

$$\delta_{\max,\breve{L},L}(\varepsilon, \mathrm{Q}) = \varepsilon^p \,, \qquad\qquad \varepsilon \geq 0.$$

Since $y_1 \neq y_2$, we then see that the assumptions of Corollary 5.37 are satisfied, and hence we have $\mathcal{R}^*_{L,\mathrm{P},F} = \mathcal{R}^*_{L,\mathrm{P}}$ if and only if F is dense in $L_p(\mathrm{P}_X)$. Moreover, it also shows that restricting the class of distributions to noise-free distributions P, i.e., to distributions with $\mathrm{P}(\,\cdot\,|x) = \delta_{\{g(x)\}}$ for measurable $g : X \to \mathbb{R}$, does not change this characterization. ◁

Example 5.39. Let $\epsilon > 0$ and L be the **ϵ-insensitive loss** defined in Example 2.42. Moreover, for $y_1 \neq y_2$, we define $\mathrm{Q}_i := \frac{1}{2}\delta_{\{y_i - \epsilon\}} + \frac{1}{2}\delta_{\{y_i + \epsilon\}}$, $i = 1, 2$. Then L, Q_1, and Q_2 satisfy the assumptions of Corollary 5.37 for $p = 1$.

In order to see this, we first recall with Example 2.42 that L is a Nemitski loss of order 1. Let us define $\psi(r) := \max\{0, |r| - \epsilon\}$, $r \in \mathbb{R}$. Then we have

$$2\mathcal{C}_{L,\mathrm{Q}_i}(t) = \psi(y_i - \epsilon - t) + \psi(y_i + \epsilon - t) \,, \qquad t \in \mathbb{R} \,, \, i = 1, 2,$$

and thus we have $\mathcal{C}_{L,\mathrm{Q}_i}(y_i) = 0 \leq \mathcal{C}_{L,\mathrm{Q}_i}(t)$ for all $t \in \mathbb{R}$. For $t \geq 0$, this yields

$$\mathcal{C}_{L,\mathrm{Q}_i}(y_i \pm t) - \mathcal{C}^*_{L,\mathrm{Q}_i} = \frac{1}{2}\psi(\epsilon + t) + \frac{1}{2}\psi(\epsilon - t) \geq \frac{1}{2}\psi(\epsilon + t) = \frac{t}{2} \,,$$

and hence we find $\mathcal{M}_{L,\mathrm{Q}_i}(0^+) = \{y_i\}$ and $\delta_{\max,\breve{L},L}(\varepsilon, \mathrm{Q}) \geq \varepsilon/2$ for $\varepsilon \geq 0$. ◁

The following example shows a similar result for the hinge loss, which is used as a surrogate for the classification loss.

Example 5.40. Let L be the **hinge loss** defined in Example 2.27. Furthermore, let Q_1, Q_2 be distributions on $Y := \{-1, 1\}$ with $\eta_1 := \mathrm{Q}_1(\{1\}) \in (0, 1/2)$ and $\eta_2 := \mathrm{Q}_2(\{1\}) \in (1/2, 1)$. Then L, Q_1, and Q_2 satisfy the assumptions of Corollary 5.37 for $p = 1$.

In order to see this, recall that L_{hinge} is a Lipschitz continuous loss, and hence it is a Nemitski loss of order 1 by (2.11). Moreover, Example 3.7 shows that $\mathcal{M}_{L_{\mathrm{hinge}},\eta}(0^+) = \{\mathrm{sign}(2\eta - 1)\}$ for $\eta \neq 0, \frac{1}{2}, 1$, and Exercise 3.15 shows that

$$\delta_{\max, L_{\mathrm{hinge}}}(\varepsilon, \eta) = \varepsilon \, \min\{\eta, 1 - \eta, 2\eta - 1\}, \qquad \varepsilon \geq 0. \qquad \lhd$$

Note that, unlike the distributions in Example 5.38, the distributions in Example 5.40 are noisy. This is due to the fact that only noise makes the hinge loss minimizer unique (see Example 3.7 and Figure 3.1 on page 54). Moreover, note that using, e.g., the least squares loss for a classification problem requires $L_2(\mu)$-denseness, which in general is a strictly stronger condition than the $L_1(\mu)$-denseness, which is sufficient for the hinge loss and the logistic loss for classification. This is somewhat remarkable since the target functions for the former two losses are bounded, whereas in general the target function for the logistic loss for classification is not even integrable.

We saw in Theorem 5.30 that "strict positive definiteness of k for infinite sequences" is *sufficient* for $\mathcal{R}^*_{L,\mathrm{P},H} = \mathcal{R}^*_{L,\mathrm{P}}$. On the other hand, "strict positive definiteness for finite sequences" is *necessary* for $\mathcal{R}^*_{L,\mathrm{P},H} = \mathcal{R}^*_{L,\mathrm{P}}$ by Corollary 5.34, and hence it seems natural to ask whether the latter, weaker condition is also sufficient. However, with the developed theory, it is easy to check that in general this is not the case. Indeed, recall that we saw in Theorem 4.62 that there exists a strictly positive definite kernel whose RKHS H is not dense in the spaces $L_1(\mu)$. Consequently, this RKHS cannot satisfy $\mathcal{R}^*_{L,\mathrm{P},H} = \mathcal{R}^*_{L,\mathrm{P}}$ for the loss functions considered in Examples 5.38 to 5.40.

Let us finally present an example of a set of bounded continuous functions that drastically fails to have good universal approximation properties.

Proposition 5.41. *Let* $Y := \{-1, 1\}$ *and* $\mathcal{P}^d_n$ *be the set of all polynomials on* $X := [0, 1]^d$ *whose degree is less than* $n + 1$*. Then, for all* $\varepsilon > 0$*, there exists a distribution* P *on* $X \times Y$ *with* $\mathcal{R}^*_{L_{\mathrm{class}},\mathrm{P}} = 0$ *and*

$$\mathcal{R}^*_{L_{\mathrm{class}},\mathrm{P},\mathcal{P}^d_n} \ \geq \ \frac{1}{2} - \varepsilon \ .$$

Moreover, P *can be chosen such that the "classes"* $\{x \in X : \mathrm{P}(y = 1|x) = 0\}$ *and* $\{x \in X : \mathrm{P}(y = 1|x) = 1\}$ *have strictly positive distance.*

Proof. We first treat the case $d = 1$. To this end, we fix an integer $m \geq (3n + 2)/\varepsilon$ and write

$$I_i := [(i+\varepsilon)/m, (i+1-\varepsilon)/m]\,, \qquad\qquad i = 0, \dots, m-1.$$

Moreover, let μ_{I_i} be the Lebesgue measure on I_i and let P be defined by

$$\mathrm{P}_X := (1-2\varepsilon)^{-1} \sum_{i=0}^{m-1} \mu_{I_i}\,,$$

$$\mathrm{P}\big(y=1\big|x\big) := \begin{cases} 1 & \text{if } x \in I_i \text{ for some even } i \in \{0, \dots, m-1\} \\ 0 & \text{otherwise.} \end{cases}$$

For a fixed polynomial $f \in \mathcal{P}_n^1$, we now write $x_1 < \ldots < x_k$, $k \leq n$ for its mutually different and ordered zeros in $(0,1)$. In addition, we write $x_0 := 0$ and $x_{k+1} := 1$. Finally, we define $a_j := |\{i : I_i \subset [x_j, x_{j+1}]\}|$ for $j = 0, \dots, k$. Obviously, there are at most k intervals I_i that do not lie between two consecutive zeros, and hence we get

$$\sum_{j=0}^{k} a_j \;\geq\; m-k \;\geq\; m-n\,.$$

Moreover, at most $\lfloor (a_j+1)/2 \rfloor$ intervals I_i are correctly classified on $[x_j, x_{j+1}]$ by the function sign$\circ f$, and consequently at least

$$\sum_{j=0}^{k} \Big\lfloor \frac{a_j}{2} \Big\rfloor \geq \sum_{j=0}^{k} \Big(\frac{a_j}{2} - 1 \Big) \geq \frac{m-n}{2} - (k+1) \geq \frac{m-3n-2}{2}$$

intervals I_i are not correctly classified on $[0,1]$ by sign$\circ f$. Since $\mathrm{P}_X(I_i) = 1/m$, we thus obtain

$$\mathcal{R}_{L_{\mathrm{class}},\mathrm{P}}(f) \;\geq\; \frac{1}{m} \sum_{j=0}^{k} \Big\lfloor \frac{a_j}{2} \Big\rfloor \;\geq\; \frac{1}{2} - \frac{3n+2}{m} \;\geq\; \frac{1}{2} - \varepsilon\,.$$

Finally, for the case $d > 1$, let $I : [0,1] \to [0,1]^d$ be the embedding defined by $t \mapsto (t, 0, \dots, 0)$, $t \in \mathbb{R}$. Moreover, consider the above distribution P embedded into $[0,1]^d$ via I. Then observe that, given an $f \in \mathcal{P}_n^d$, its restriction $f_{|I([0,1])}$ can be interpreted as a polynomial in $\mathcal{P}_n^1$, and from this it is straightforward to prove the assertion. □

5.6 Further Reading and Advanced Topics

The first representer theorem for empirical distributions was established for some specific losses, including the least squares loss, by Kimeldorf and Wahba (1971). The version we presented in Theorem 5.5 is a simplified version of a more general representer theorem for empirical distributions proved by

Schölkopf *et al.* (2001b). For a brief discussion on some related results, we refer to these authors as well as to the book by Schölkopf and Smola (2002). Finally, Dinuzzo *et al.* (2007) recently established a refinement of Theorem 5.8 for empirical distributions.

The first considerations on general SVM solutions were made by Zhang (2001). In particular, he showed that the general SVM solutions exist for the hinge loss and that they converge to the hard margin SVM solution for $\lambda \to 0$ (see Exercise 5.10 for a precise statement). Moreover, he also mentions the general representer theorem for the hinge loss, though he does not prove it. The existence of general SVM solutions for a wide class of L_{class}-surrogates was then established by Steinwart (2005), and a corresponding representer theorem was established by Steinwart (2003). The existence of general SVM solutions for convex integrable Nemitski losses of some type $p \geq 1$ as well as the corresponding general representer theorem was then shown by De Vito *et al.* (2004). The results presented in Section 5.2 are closely related to their findings, although their representer theorem is stated for supervised losses and for SVM optimization problems having an additional "offset".

The Lipschitz continuity of general SVM solutions we presented in Section 5.3 was again found by Zhang (2001) for differentiable losses. Independently, Bousquet and Elisseeff (2002) established a similar result for empirical SVM solutions (see Exercise 5.6). The presentation given in Section 5.3 mainly follows that of Christmann and Steinwart (2007).

The relation between the approximation error functions A_2 and A_∞ was investigated by Steinwart and Scovel (2005a). The presentation of the corresponding part of Section 5.4 closely follows their work, though it is fair to say that some of the results, such as Lemma 5.23, are to some extent folklore. Furthermore, it is interesting to note that the functions A_2 and A_∞ are often closely related to concepts from approximation theory. To be more precise, let E and F be Banach spaces such that $E \subset F$ and $\mathrm{id} : E \to F$ is continuous. Then the **K-functional** is defined by

$$K(y,t) := \inf_{x \in E} t\|x\|_E + \|y - x\|_F\,, \qquad\qquad y \in F,\, t > 0.$$

Moreover, for $r \in (0,1)$, the **interpolation space** $(E,F)_r$ is the set of elements $y \in F$ for which

$$\|y\|_{(E,F)_r} := \sup_{t>0} K(y,t)t^{-r} < \infty.$$

One can show that $((E,F)_r, \|\cdot\|_{(E,F)_r})$ is a Banach space, and for many classical spaces E and F this interpolation space can actually be described in closed form. We refer to the books by Bergh and Löfström (1976), Triebel (1978), and Bennett and Sharpley (1988). In particular, for a Euclidean ball $X \subset \mathbb{R}^d$, Theorem 7.31 of Adams and Fournier (2003) shows that the Sobolev space $W^k(X)$ is continuously embedded into $(W^m(X), L_2(X))_{k/m}$ for all $0 < k < m$ and that this embedding is almost sharp. Besides the K-functional, one can also consider the functional

$$A(y,t) := \inf_{x \in tB_E} \|x-y\|_F\,, \qquad\qquad y \in F,\, t > 0.$$

Not surprisingly, both functionals are closely related. Indeed, with techniques similar to those of Theorem 5.25, Smale and Zhou (2003) showed that

$$y \in (E,F)_r \qquad\qquad \Longleftrightarrow \qquad\qquad c := \sup_{t>0} A(y,t) t^{\frac{r}{1-r}} < \infty\,,$$

and in this case we actually have $c^{1-r} \leq \|y\|_{(E,F)_r} \leq 2c^{1-r}$. Let us now illustrate how these abstract results relate to the approximation error functions. To this end, let us first assume that L is the least squares loss. Moreover, we fix an RKHS H over X with bounded measurable kernel and a distribution P on $X \times \mathbb{R}$ with $|\mathrm{P}|_2 < \infty$. Then we have $\mathcal{R}_{L,\mathrm{P}}(f) - \mathcal{R}^*_{L,\mathrm{P}} = \|f - f^*_{L,\mathrm{P}}\|^2_{L_2(\mathrm{P}_X)}$, and by setting $E := H$ and $F := L_2(\mathrm{P}_X)$, we thus find $A_\infty(t) = A^2(f^*_{L,\mathrm{P}}, t^{-1})$ for all $t > 0$. Using the relation between A_∞ and A_2, we conclude that

$$f^*_{L,\mathrm{P}} \in \bigl(H, L_2(\mathrm{P}_X)\bigr)_r \qquad\qquad \Longleftrightarrow \qquad\qquad \exists\, c \geq 0\ \forall \lambda > 0 : A_2(\lambda) \leq c\lambda^r\,,$$

and if $f^*_{L,\mathrm{P}} \in (H, L_2(\mathrm{P}_X))_r$ we may actually choose $c := \|f^*_{L,\mathrm{P}}\|^2_{(H,L_2(\mathrm{P}_X))_r}$. In particular, if $H = W^m(X)$, $f^*_{L,\mathrm{P}} \in W^k(X)$ for some $k < m$, and P_X is the uniform distribution on X, then there exists a constant $c > 0$ that is independent of $f^*_{L,\mathrm{P}}$ such that

$$A_2(\lambda) \leq c\, \|f^*_{L,\mathrm{P}}\|^2_{W^k(X)} \lambda^{k/m}\,, \qquad\qquad \lambda > 0\,.$$

In this direction, it is also interesting to note that Smale and Zhou (2003) showed that, given a *fixed* width $\gamma > 0$, the approximation error function of the corresponding Gaussian RBF kernel can only satisfy a non-trivial bound of the form $A_2(\lambda) \leq c\lambda^\beta$ if $f^*_{L,\mathrm{P}}$ is C^∞. Furthermore, for the least squares loss, the approximation error function is also related to the integral operator of the kernel of H. To explain this, let X be a compact metric space and $T_k : L_2(\mathrm{P}_X) \to L_2(\mathrm{P}_X)$ be the integral operator associated to the kernel k and the measure P_X. Then one can show (see, e.g., Theorem 4.1 in Cucker and Zhou, 2007) that

$$f^*_{L,\mathrm{P}} \in T_k^{r/2}(L_2(\mathrm{P}_X)) \qquad\qquad \Longrightarrow \qquad\qquad \exists\, c \geq 0\ \forall \lambda > 0 : A_2(\lambda) \leq c\lambda^r\,.$$

In addition, this theorem also shows that if the latter estimate on A_2 holds and $\operatorname{supp} \mathrm{P}_X = X$, then we have $f^*_{L,\mathrm{P}} \in T_k^{(r-\varepsilon)/2}(L_2(\mathrm{P}_X))$ for all $\varepsilon > 0$. Finally, let us assume that L is a Lipschitz continuous loss for which a Bayes decision function $f^*_{L,\mathrm{P}}$ exists. Then we have $\mathcal{R}_{L,\mathrm{P}}(f) - \mathcal{R}^*_{L,\mathrm{P}} \leq \|f - f^*_{L,\mathrm{P}}\|_{L_1(\mathrm{P}_X)}$, and by setting $E := H$ and $F := L_1(\mathrm{P}_X)$ we thus find $A_\infty(t) \leq A(f^*_{L,\mathrm{P}}, t^{-1})$ for all $t > 0$. From this we conclude that

$$f^*_{L,\mathrm{P}} \in \bigl(H, L_1(\mathrm{P}_X)\bigr)_r \qquad \Longrightarrow \qquad \forall \lambda > 0 : A_2(\lambda) \leq \|f^*_{L,\mathrm{P}}\|^{\frac{2}{2-r}}_{(H,L_1(\mathrm{P}_X))_r} \lambda^{\frac{r}{2-r}}\,.$$

Qualitative approximation properties of function classes in terms of excess risks have been investigated for quite a while. For example, certain neural network architectures have been known for more than 15 years to be universal approximators. For some information in this regard, we refer to p. 518f of the book by Devroye *et al.* (1996). Moreover, the characterization for the least squares loss given in Example 5.38 is, of course, trivial if we recall the formula for the excess least squares risk given in Example 2.6. Moreover, Example 5.38 can also be shown without using the self-calibration function. Furthermore, the fact that $L_\infty(\mu)$-denseness is sufficient for most commonly used loss functions is also to some extent folklore. The more sophisticated sufficient condition given in Theorem 5.31 together with the general necessary conditions found in Section 5.5 are taken from Steinwart *et al.* (2006b). Finally, Corollary 5.29 together with the universality of certain kernels was first shown by Steinwart (2001, 2005) and independently by Hammer and Gersmann (2003).

5.7 Summary

In this chapter, we investigated general SVM solutions and their properties. To this end, we showed in the first section that for a large class of convex losses these solutions exist and are unique. Being motivated by the representer theorem for empirical solutions, we then established a general representer theorem, which additionally describes the form of the representing function, in the second section. In the third section, we used this general representer theorem to show that the general SVM solutions depend on the underlying distribution in a Lipschitz continuous fashion.

In the fourth section, we investigated the behavior of the general SVM solution and its associated (regularized) risk for vanishing regularization parameters. In particular, we showed that this risk tends to the best possible risk obtainable in the RKHS. In addition, we compared the regularization scheme of SVMs with a more classical notion of approximation error.

In the last section, we investigated under which assumptions the RKHS is rich enough to achieve the Bayes risk. Here we first derived a sufficient condition for universal kernels. We then obtained a characterization for a large class of loss functions and illustrated these findings with some examples.

5.8 Exercises

5.1. Existence of general SVM solutions for pinball loss ($\star$)

Formulate a condition on P that ensures the existence of a general SVM solution when using the pinball loss. Does this condition change for different values of τ?

5.2. A representer theorem for the logistic loss ($\star\star$)
Consider a strictly positive definite kernel and the logistic loss for classification. Show that all coefficients in (5.6) do not vanish.

5.3. A representer theorem for some distance-based losses ($\star\star$)
Formulate a generalized representer theorem for the loss functions $L_{p\text{-dist}}$, $p \geq 1$, and $L_{\text{r-logist}}$. Compare the different norm bounds for the representing function h. What does (5.10) mean for the least squares loss?

5.4. Generalized representer theorem for margin-based losses ($\star\star$)
Formulate a generalized representer theorem for convex, margin-based loss functions. How can (5.19) and (5.21) be simplified?

5.5. Generalized representer theorem for the hinge loss ($\star\star\star$)
Formulate a generalized representer theorem for the hinge loss. What is the form of (5.19) and (5.21)? Then assume that P is an empirical distribution with respect to a data set $D = ((x_1, y_1), \dots, (x_n, y_n)) \in (X \times Y)^n$. Investigate the form of the coefficients in (5.6) and describe when a specific coefficient vanishes. Finally, find a condition on the kernel that ensures that the coefficients are uniquely determined.

5.6. Stability of empirical SVM solutions ($\star\star$)
Let $D := ((x_1, y_1), \dots, (x_n, y_n)) \in (X \times Y)^n$ be a sample set and $\bar{D} := ((x_1, y_1), \dots, (x_{n-1}, y_{n-1})) \in (X \times Y)^{n-1}$ be the sample set we obtain from D by removing the last sample. Furthermore, let $L : X \times Y \times \mathbb{R} \to [0, \infty)$ be a convex loss function, H be a RKHS over X with canonical feature map $\Phi : X \to H$, and $\lambda > 0$. Show that

$$\| f_{\mathrm{D},\lambda} - f_{\bar{\mathrm{D}},\lambda} \|_H \leq \frac{2\|k\|_\infty}{\lambda n} \cdot |L|_{B_\lambda,1} ,$$

where B_λ is defined by (5.29).

5.7. Existence of general SVM solutions for non-convex losses ($\star\star\star\star$)
Let $L : X \times Y \times \mathbb{R} \to [0, \infty)$ be a continuous loss function and P be a distribution on $X \times Y$ such that L is a P-integrable Nemitski loss. Furthermore, let H be the RKHS of a bounded measurable kernel over X. Show that for all $\lambda > 0$ there exists a general SVM solution $f_{\mathrm{P},\lambda}$.

Hint: Adapt the technique used in the proof of Theorem 5.17.

5.8. Approximation error function without RKHSs ($\star\star\star$)
Investigate which results from Section 5.4 still hold if the RKHS H in the definition of the approximation error functions is replaced by a general normed space consisting of measurable functions $f : X \to \mathbb{R}$.

5.9. Approximation error function with general exponent ($\star\star\star\star$)
Let $L : X \times Y \times \mathbb{R} \to [0, \infty)$ be a loss function, H be the RKHS of a measurable

kernel over X, and P be a distribution on $X \times Y$ with $\mathcal{R}^*_{L,\mathrm{P},H} < \infty$. For $p \geq 1$, define the *p-approximation error function* $A_p : [0,\infty) \to [0,\infty)$ by

$$A_p(\lambda) := \inf_{f \in H} \lambda \|f\|_H^p + \mathcal{R}_{L,\mathrm{P}}(f) - \mathcal{R}^*_{L,\mathrm{P},H}\,, \qquad \lambda \geq 0.$$

Establish a (suitably modified) version of Lemma 5.15. Furthermore, show the uniqueness and existence of a minimizer of $A_p(\lambda)$ under the assumptions of Lemma 5.1 and Theorem 5.2. Then establish (suitably modified) versions of Theorem 5.17 and its Corollaries 5.18, 5.19, and 5.24. Finally, discuss the consequences of the latter with respect to variable exponent p.

5.10. SVM behavior for classes with strictly positive distances ($\star\star$)
Let X be a compact metric space, $Y := \{-1, 1\}$, and P be a distribution on $X \times Y$ with $\mathcal{R}^*_{L_{\mathrm{class}},\mathrm{P}} = 0$. Let us write $\eta(x) := \mathrm{P}(y = 1|x)$, $x \in X$. Assume that the "classes"

$$\{x \in X : \eta = 0\} \cap \operatorname{supp} \mathrm{P}_X$$

and

$$\{x \in X : \eta = 1\} \cap \operatorname{supp} \mathrm{P}_X$$

have strictly positive distance and that H is the RKHS of a universal kernel on X. Show that $f^*_{L_{\mathrm{hinge}},\mathrm{P},H}$ exists and that the general SVM solutions with respect to the hinge loss satisfy $\lim_{\lambda \to 0^+} f_{L_{\mathrm{hinge}},\mathrm{P},H,\lambda} = f^*_{L_{\mathrm{hinge}},\mathrm{P},H}$.

5.11. Same behavior of different approximation functions ($\star\star\star$)
Let $L : X \times Y \times \mathbb{R} \to [0,\infty)$ be a loss function, H be the RKHS of a measurable kernel over X, and P be a distribution on $X \times Y$ with $\mathcal{R}^*_{L,\mathrm{P},H} < \infty$. Moreover, for $p \geq 1$, define the *p-approximation error function* $A_p : [0,\infty) \to [0,\infty)$ as in Exercise 5.9. Finally, let $c > 0$, $\lambda_0 > 0$, and $\alpha > 0$ be fixed constants. Show the following statements:

i) $A_p(\lambda) \leq c^{\frac{p}{\alpha+p}} \lambda^{\frac{\alpha}{\alpha+p}}$ for all $\lambda > 0$ implies $A_\infty(\lambda) \leq c\lambda^\alpha$ for all $\lambda > 0$.

ii) $A_\infty(\lambda) \leq c\lambda^\alpha$ for all $\lambda > 0$ implies $A_p(\lambda) \leq 2c^{\frac{p}{\alpha+p}} \lambda^{\frac{\alpha}{\alpha+p}}$ for all $\lambda > 0$.

iii) $A_p(\lambda) \geq 2c^{\frac{p}{\alpha+p}} \lambda^{\frac{\alpha}{\alpha+p}}$ for all $\lambda \in [0, \lambda_0]$ implies $A_\infty(\lambda) \geq c\lambda^\alpha$ for all $\lambda > 0$ with $\lambda^p A_\infty(\lambda) \leq \lambda_0$.

iv) $A_\infty(\lambda) \geq c\lambda^\alpha$ for all $\lambda \in [0, \lambda_0]$ implies $A_p(\lambda) \geq c^{\frac{p}{\alpha+p}} \lambda^{\frac{\alpha}{\alpha+p}}$ for all $\lambda > 0$ with $\frac{\lambda}{A_p(\lambda)} \leq \lambda_0^p$.

With the help of these estimates, discuss the optimality of (5.41) for $p = 2$.

Hint: First prove an analogue to Theorem 5.25.

5.12. Some other conditions for universal approximators ($\star\star$)
Show that Theorem 5.28 still holds true if we only assume that for all $g \in L_\infty(\mathrm{P}_X)$ there exists a sequence $(f_n) \subset F$ with $\sup_{n \geq 1} \|f_n\|_\infty < \infty$ and $\lim_{n \to \infty} f_n = g$ in probability P_X.

5.13. Universal approximators for some margin-based losses ($\star\star$)
Discuss the squared hinge loss and the least squares loss in the sense of Example 5.40.