8

Support Vector Machines for Classification

Overview. *Classification is one of the main areas of application for support vector machines. This chapter presents key results on the generalization performance of SVMs when applied to this learning problem. In addition, we investigate how the choice of the loss influences both the number of support vectors we can typically expect and the ability of SVMs to estimate posterior probabilities.*

Prerequisites. *For the basic results on the generalization performance, we need Sections 2.1–2.3 on loss functions, Chapter 4 on kernels, Chapter 5 on infinite-sample SVMs, and the oracle inequalities from Chapter 6. The more advanced results in this direction also require Chapter 7. For estimating the number of support vectors and the posterior probabilities, we further need Sections 3.1–3.6 and 3.9.*

Usage. *Parts of this chapter are helpful in Section 10.3 on robustness of SVMs for classification and in Chapter 11, where practical aspects of SVMs are discussed.*

Binary classification is one of the central applications for machine learning methods in general and for SVMs in particular. In this chapter, we investigate features of SVMs when applied to binary classification. In particular, we consider the following questions:

- For what types of distributions do SVMs learn fast?
- How many support vectors can we expect?
- Can SVMs be used to estimate posterior probabilities?
- Which loss functions are a reasonable alternative to the hinge loss?

Obviously, the first question is *the* main question since the primary goal in almost every application dealing with classification is a good classification performance. However, in many applications, other features are also highly desired such as *a)* a low computational complexity in the training and/or the employment phase and *b)* the ability to estimate the probability $\eta(x)$ of a positive label y at point x. Now it is obvious that the number of support vectors has a direct influence on the time required to evaluate the SVM decision function, and therefore the second question addresses the computational complexity during the employment phase. Moreover, we will see in Chapter 11 that the number of support vectors also has a substantial influence on the time required to train the SVM, and therefore the second question actually

addresses the overall computational complexity. Finally, though the hinge loss is the loss function that is most often used in practice, it has some disadvantages. First, it is not differentiable, and hence certain optimization procedures cannot be applied to its SVM optimization problem. Second, we have seen in Chapters 2 and 3 that the minimizer of the hinge loss is $\operatorname{sign}(2\eta(x)-1)$, $x \in X$. However, this expression does not contain information about the size of $\eta(x)$, and therefore SVM decision functions obtained by using the hinge loss cannot be used to estimate $\eta(x)$. This discussion shows that a deliberative decision on whether or which SVM is used for a particular application requires at least answers to the questions above. Moreover, particular applications may require taking into account further aspects of learning methods such as robustness, considered in Section 10.3.

The rest of this chapter is organized as follows. In Section 8.1, we reformulate some basic oracle inequalities from Chapter 6. These oracle inequalities estimate the excess *classification* risk of SVMs using the hinge loss and can be used to develop data-dependent parameter selection strategies such as the one considered in Section 6.5. In Section 8.2, we will then focus on SVMs that use a Gaussian RBF kernel since these are the kernels that are most often used in practice. For such SVMs, we present a rather general assumption on the data-generating distribution P that describes the behavior of P in the vicinity of the decision boundary and that enables us to estimate the approximation error function. This is then used to analyze a training validation support vector machine (TV-SVM) that uses the validation set to determine both the regularization parameter and the kernel parameter. Section 8.3 then uses the more advanced oracle inequalities of Chapter 7 to improve the learning rates obtained in Section 8.2. To this end, another assumption on P is introduced that measures the amount of noise P has in the labeling process and that can be used to establish a variance bound for the hinge loss. In Section 8.4, we then investigate the sparseness of SVMs by presenting a lower bound on the number of support vectors. Here it will turn out that the Bayes classification risk is the key quantity that asymptotically controls the sparseness for SVMs using the hinge loss. In the last section, we will then consider SVMs for binary classification that use an alternative loss function. Here we first improve the general calibration inequalities obtained in Section 3.4 for distributions satisfying the low-noise assumption introduced in Section 8.3. Furthermore, we will establish a general lower bound on the sparseness of SVMs that use a margin-based loss. Finally, this lower bound is used to describe some properties of loss functions that prevent the decision functions from being sparse.

8.1 Basic Oracle Inequalities for Classifying with SVMs

The goal of this section is to illustrate how the basic oracle inequalities established in Section 6.4 can be used to investigate the classification performance of SVMs using the hinge loss.

Given a distribution P on $X \times Y$, we assume throughout this and the following sections that P^n denotes the canonical extension of n-fold product measure of P to the universal completion of the product σ-algebra on $(X \times Y)^n$. As in Section 6.4, this assumption makes it possible to ignore measurability questions.

Let us now begin with a reformulation of Theorem 6.24.

Theorem 8.1 (Oracle inequality for classification). *Let L be the hinge loss, $Y := \{-1, 1\}$, H be a separable RKHS with bounded measurable kernel k over X satisfying $\|k\|_\infty \leq 1$, and P be a distribution on $X \times Y$ such that H is dense in $L_1(\mathrm{P}_X)$. Then, for all $\lambda > 0$, $n \geq 1$, and $\tau > 0$, we have with probability P^n not less than $1 - e^{-\tau}$ that*

$$\mathcal{R}_{L_{\text{class}},\mathrm{P}}(f_{\mathrm{D},\lambda}) - \mathcal{R}^*_{L_{\text{class}},\mathrm{P}} < A_2(\lambda) + \lambda^{-1}\left(\sqrt{\frac{8\tau}{n}} + \sqrt{\frac{4}{n}} + \frac{8\tau}{3n}\right),$$

where $A_2(\,\cdot\,)$ is the approximation error function with respect to L, H, and P.

Proof. Obviously, $L := L_{\text{hinge}}$ is a convex and Lipschitz continuous loss satisfying $L(y, 0) = 1$ for all $y \in Y$. Therefore, Theorem 6.24 shows that

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

holds with probability P^n not less than $1 - e^{-\tau}$. Moreover, the hinge loss is also a P-integrable Nemitski loss of order 1 by Lemma 2.25, and hence Theorem 5.31 yields $\mathcal{R}^*_{L,\mathrm{P},H} = \mathcal{R}^*_{L,\mathrm{P}}$. Consequently, we obtain by Theorem 2.31 that

$$\mathcal{R}_{L_{\text{class}},\mathrm{P}}(f_{\mathrm{D},\lambda}) - \mathcal{R}^*_{L_{\text{class}},\mathrm{P}} \;\leq\; \mathcal{R}_{L,\mathrm{P}}(f_{\mathrm{D},\lambda}) - \mathcal{R}^*_{L,\mathrm{P},H}\,. \qquad \square$$

Note that the left-hand side of the oracle inequality above considers the excess *classification* risk rather than the excess hinge risk, while the SVM is assumed to use the hinge loss. Consequently, if we can ensure that the right-hand side of the oracle inequality converges to 0, then the SVM is consistent with respect to the classification risk, and this finally justifies the use of the hinge loss as a surrogate for the classification loss.

We have seen in Section 6.4 that the oracle inequality above can sometimes be improved if the unit ball of the RKHS used has finite $\|\cdot\|_\infty$-covering numbers. The following theorem reformulates this result for polynomially growing covering numbers, i.e., polynomially decaying entropy numbers.

Theorem 8.2 (Classification with benign kernels). *Let L be the hinge loss, $Y := \{-1, 1\}$, X be a compact metric space, and H be the RKHS of a continuous kernel k over X with $\|k\|_\infty \leq 1$. Moreover, assume that there exist constants $a \geq 1$ and $p \in (0, 1]$ such that the dyadic entropy numbers satisfy*

$$e_i\big(\mathrm{id} : H \to C(X)\big) \;\leq\; a\, i^{-\frac{1}{2p}}\,, \qquad\qquad i \geq 1. \tag{8.1}$$

Then, for all distributions P *on* $X \times Y$*, for which* H *is dense in* $L_1(\mathrm{P}_X)$*, and all* $\lambda \in (0,1]$*,* $n \geq 1$*,* $\tau > 0$*, we have*

$$\mathcal{R}_{L_{\text{class}},\mathrm{P}}(f_{\mathrm{D},\lambda}) - \mathcal{R}^*_{L_{\text{class}},\mathrm{P}} < A_2(\lambda) + 8\Big(\frac{a^{2p}}{\lambda^{1+p}\, n}\Big)^{1/(2+2p)} + \sqrt{\frac{8\tau+8}{\lambda n}}$$

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

Proof. Let us fix an $\varepsilon > 0$. Using Theorem 6.25, we see analogously to the proof of Theorem 8.1 that

$$\mathcal{R}_{L_{\text{class}},\mathrm{P}}(f_{\mathrm{D},\lambda}) - \mathcal{R}^*_{L_{\text{class}},\mathrm{P}} < A_2(\lambda) + 4\varepsilon + \sqrt{\frac{8\tau+8\ln\big(2\mathcal{N}(B_H, \|\cdot\|_\infty, \lambda^{\frac{1}{2}}\varepsilon)\big)}{\lambda n}}$$

holds with probability P^n not less than $1 - e^{-\tau}$. Moreover, (8.1) implies

$$\ln \mathcal{N}(B_H, \|\cdot\|_\infty, \varepsilon) \;\leq\; \ln(4) \cdot \Big(\frac{a}{\varepsilon}\Big)^{2p}, \qquad\qquad \varepsilon > 0\,,$$

by Lemma 6.21. Combining these estimates, optimizing the result with respect to ε by Lemma A.1.5, and using $(1+p)(4/p)^{p/(1+p)}(8\ln 4)^{1/(2+2p)} \leq 8$ then yields the assertion. □

Since the hinge loss is Lipschitz continuous, it is straightforward to check that the oracle inequalities above produce the learning rates we have discovered at the end of Section 6.4 and in Exercise 6.9. However, in this case, the learning rates are for the classification risk instead of the hinge risk. Moreover, note that we can easily use the oracle inequalities above to derive learning rates for data-dependent parameter selection strategies such as the TV-SVM. Since the oracle inequalities above consider the classification risk, we can, however, use either the clipped empirical hinge risk *or* the empirical classification risk in the validation step. We come back to this observation at the end of the following section, where we investigate a modified TV-SVM that also selects a kernel parameter in a data-dependent fashion.

8.2 Classifying with SVMs Using Gaussian Kernels

The Gaussian RBF kernel is one of the most often used kernels in practice. In this section, we derive learning rates for situations in which this kernel is used with different parameter values. To this end, we introduce some conditions on the data-generating distribution that describe their behavior near the "decision boundary". For such distributions, we then establish a bound on the approximation error function, which in turn will yield the learning rates. Finally, we discuss a method to select both the regularization parameter and the kernel parameter in a data-dependent, adaptive way.

Let us begin by presenting an oracle inequality for SVMs using the hinge loss and a Gaussian kernel.

Theorem 8.3 (Oracle inequality for Gaussian kernels). *Let L be the hinge loss, $Y := \{-1,1\}$, $X \subset \mathbb{R}^d$ be a compact subset, and $m \in \mathbb{N}$. Then there exists a constant $c_{m,d}(X) \geq 1$ such that, for all distributions P on $X \times Y$ and all fixed $\gamma, \lambda \in (0,1]$, $n \geq 1$, $\tau > 0$, we have with probability P^n not less than $1 - e^{-\tau}$ that*

$$\mathcal{R}_{L_{\text{class}},\mathrm{P}}(f_{\mathrm{D},\lambda,\gamma}) - \mathcal{R}^*_{L_{\text{class}},\mathrm{P}} < A_2^{(\gamma)}(\lambda) + c_{m,d}(X)\lambda^{-1/2}\bigl(\gamma^d\, n\bigr)^{-\frac{m}{2m+d}}\sqrt{\tau+1}\,.$$

Here $A_2^{(\gamma)}$ denotes the approximation error function with respect to L, P, and the Gaussian RBF kernel k_γ, and $f_{\mathrm{D},\lambda,\gamma}$ denotes the decision function of an SVM using L and k_γ.

Proof. Since X is compact, we may assume without loss of generality that X is a closed Euclidean ball. Then the assertion follows from combining Theorem 6.27 with Theorem 8.2. □

With some extra effort one can, in principle, determine a value for the constant $c_{m,d}(X)$. However, this requires us to find a constant in a bound on entropy numbers and hence we omit the details. Besides the constant $c_{m,d}(X)$, we also need to bound the approximation error function $A_2^{(\gamma)}$ in order to obtain learning rates from Theorem 8.3. Unfortunately, this requires us to impose assumptions on P since otherwise Theorem 8.3 would provide us with a uniform learning rate for classification, which by Corollary 6.7 is impossible.

In order to formulate such a condition, let us recall that "the" regular conditional probability $\mathrm{P}(\,\cdot\,|x)$ of a probability measure P on $X \times \{-1,1\}$ is only P_X-almost surely defined by (A.11). In other words, if we have two measurable functions $\eta_1, \eta_2 : X \to [0,1]$ for which the probability measures $\mathrm{P}_i(\,\cdot\,|x)$ defined by $\mathrm{P}_i(y=1|x) := \eta_i(x)$, $i = 1,2$, $x \in X$, satisfy (A.11), then η_1 and η_2 coincide up to a P_X-zero set. Conversely, any measurable modification of, say, η_1, on a P_X-zero set yields a regular conditional probability of P. Since in the following considerations we have to deal with this inherent ambiguity very carefully, we introduce the following definition.

Definition 8.4. *Let $Y := \{-1,1\}$, X be a measurable space, and P be a distribution on $X \times Y$. We say that a measurable function $\eta : X \to [0,1]$ is a **version of the posterior probability** of P if the probability measures $\mathrm{P}(\,\cdot\,|x)$ defined by $\mathrm{P}(y=1|x) := \eta(x)$, $x \in X$, form a regular conditional probability of P, i.e.,*

$$\mathrm{P}(A \times B) = \int_A \mathrm{P}(B|x)\, d\mathrm{P}_X(x)\,,$$

for all measurable sets $A \subset X$ and $B \subset Y$.

If a distribution P on $X \times Y$ has a smooth version of the posterior probability, then intuitively the set $\{x \in X : \eta = 1/2\}$ is the "decision boundary" that separates the class $\{x \in X : \eta(x) < 1/2\}$ of negatively labeled samples

from the class $\{x \in X : \eta(x) > 1/2\}$ of positively labeled samples. Intuitively, any classification method that uses a notion of distance on X to build its decision functions should learn well in regions that are not close to the decision boundary. This suggests that learning should be relatively easy for distributions that do not have a lot of mass in the vicinity of the decision boundary. Our next goal is to verify this intuition for SVMs using Gaussian kernels. To this end, we begin with the following definition that introduces a "distance to the decision boundary" without defining the decision boundary itself.

Definition 8.5. *Let (X, d) be a metric space, P be a distribution on $X \times \{-1, 1\}$, and $\eta : X \to [0, 1]$ be a version of its posterior probability. We write*

$$X_{-1} := \bigl\{x \in X : \eta(x) < 1/2\bigr\}\,,$$
$$X_1 := \bigl\{x \in X : \eta(x) > 1/2\bigr\}\,.$$

Then the associated ***version of the distance to the decision boundary*** *is the function $\Delta : X \to [0, \infty]$ defined by*

$$\Delta(x) := \begin{cases} d(x, X_1) & \text{if } x \in X_{-1}, \\ d(x, X_{-1}) & \text{if } x \in X_1, \\ 0 & \text{otherwise}\,, \end{cases} \tag{8.2}$$

where, as usual, $d(x, A) := \inf_{x' \in A} d(x, x')$.

In the following, we are mainly interested in the distance to the decision boundary for distributions defined on $X \times \{-1, 1\}$, where $X \subset \mathbb{R}^d$ is measurable. In this case, we will always assume that these subsets are equipped with the Euclidean metric.

One may be tempted to think that changing the posterior probability on a P_X-zero set has only marginal or even no influence on Δ. Unfortunately, quite the opposite is true. To illustrate this, let us consider the probability measure P on $\mathbb{R} \times \{-1, 1\}$ whose marginal distribution P_X is the uniform distribution on $[-1, 1]$ and for which $\eta(x) := \mathbf{1}_{[0,\infty)}(x)$, $x \in \mathbb{R}$, is a version of its posterior probability. Obviously this definition gives $X_{-1} = (-\infty, 0)$ and $X_1 = [0, \infty)$, and from these equations we immediately conclude that $\Delta(x) = |x|$, $x \in \mathbb{R}$, for the associated version of the distance to the decision boundary. Let us now change the posterior probability on the P_X-zero set $\mathbb{Q}$ by defining $\tilde{\eta}(x) := \eta(x)$ if $x \in \mathbb{R} \setminus \mathbb{Q}$ and $\tilde{\eta}(x) := 1 - \eta(x)$ if $x \in \mathbb{Q}$. Since $\mathrm{P}_X(\mathbb{Q}) = 0$, we easily see that $\tilde{\eta}$ is indeed another version of the posterior probability of P. However, for this version, we have $X_{-1} = ((-\infty, 0) \setminus \mathbb{Q}) \cup ([0, \infty) \cap \mathbb{Q})$ and $X_1 = ((-\infty, 0) \cap \mathbb{Q}) \cup ([0, \infty) \setminus \mathbb{Q})$, and therefore its associated distance to the decision boundary is given by $\Delta(x) = 0$ for all $x \in \mathbb{R}$. This example[1]

[1] Informally speaking, the example exploited the fact that sets that are "small" in a measure theoretic sense can be "big" in a topological sense. From this it becomes obvious that the described phenomenon can be observed for a variety of distributions on $\mathbb{R}^d$.

demonstrates that it is absolutely necessary to fix a version of the posterior probability when dealing with Δ.

Our next goal is to use the distance to the decision boundary Δ to describe the behavior of distributions P near the decision boundary. This behavior will be crucial when bounding the approximation error function for Gaussian kernels below. Let us begin by measuring the concentration of P_X near the decision boundary. To this end, we first note that $\{x \in X : \Delta(x) < t\}$ contains the set of points x that are either *a)* in $X_{-1} \cup X_1$ and geometrically close to the opposite class or *b)* satisfy $\eta(x) = 1/2$. Consequently, in the case $P_X(\{x \in X : \eta(x) = 1/2\}) = 0$, we see that $\{x \in X : \Delta(x) < t\}$ contains only the points (modulo a P_X-zero set) that are close to the opposite class. The following definition describes the size of this set.

Definition 8.6. *Let (X, d) be a metric space and P be a distribution on $X \times \{-1, 1\}$. We say that P has* ***margin exponent*** *$\alpha \in [0, \infty)$ for the version $\eta : X \to [0, 1]$ of its posterior probability if there exists a constant $c > 0$ such that the associated version Δ of the distance to the decision boundary satisfies*

$$P_X\big(\{x \in X : \Delta(x) < t\}\big) \;\le\; c\, t^{\alpha}\,, \qquad\qquad t \ge 0. \tag{8.3}$$

In the following, we sometimes say that P has margin exponent α for the version Δ of the distance to the decision boundary if (8.3) is satisfied. Moreover, we say that P has margin exponent α if there exists a version Δ of the distance to the decision boundary such that (8.3) is satisfied.

We will see later in this section that we are mainly interested in the behavior of P_X $(\{x \in X : \Delta(x) < t\})$ for $t \to 0$. In this case, an exponent $\alpha > 0$ in (8.3) describes how *fast* $P_X(\{x \in X : \Delta(x) < t\})$ converges to 0. Now note that for $\alpha > 0$ it is straightforward to check that $P_X(\{x \in X : \Delta(x) = 0\}) = 0$ holds. Since $\eta(x) = 1/2$ implies $\Delta(x) = 0$, we hence find

$$P_X\big(\{x \in X : \eta(x) = 1/2\}\big) = 0\,.$$

Informally speaking, (8.3) therefore measures the "size" of the set of points that are close to the opposite class. We refer to Figure 8.1 for an illustration of a distribution having a large margin exponent.

Obviously, every distribution has margin exponent $\alpha = 0$. Moreover, the margin exponent is monotone in the sense that every distribution P that has some margin exponent α also has margin exponent α' for all $\alpha' \in [0, \alpha]$. The following simple yet useful lemma shows that the margin exponent is invariant with respect to inclusions of the input space X.

Lemma 8.7. *Let $(\tilde{X}, d)$ be a metric space, $X \subset \tilde{X}$ be a measurable subset, and P be a distribution on $X \times Y$, where $Y := \{-1, 1\}$. Then there exists exactly one distribution $\tilde{P}$ on $\tilde{X} \times Y$, called the* ***canonical extension*** *of P, that satisfies the following two conditions:*

i) $\tilde{P}_X(A) = P_X(A \cap X)$ for all measurable $A \subset \tilde{X}$.

ii) $\tilde{\mathrm{P}}(y = 1|x) = \mathrm{P}(y = 1|x)$ *for* P_X*-almost all* $x \in X$*.*

Moreover, P *has margin exponent* $\alpha \in [0, \infty)$ *if and only if its canonical extension* $\tilde{\mathrm{P}}$ *has margin exponent* α*.*

Proof. Obviously, *i)* can be used to define $\tilde{\mathrm{P}}_X$, and it is furthermore clear that there is only one distribution on $\tilde{X}$ that satisfies *i)*. Moreover, *i)* implies that $\tilde{X} \setminus X$ is a $\tilde{\mathrm{P}}_X$-zero set and hence the behavior of $\tilde{\mathrm{P}}(y = 1|x)$ on $\tilde{X} \setminus X$ does not matter for the definition of $\tilde{\mathrm{P}}$. By using *ii)* as a definition for the posterior probability of $\tilde{\mathrm{P}}$, we then find the first assertion, where we use (A.11) to define $\tilde{\mathrm{P}}$. In order to prove the second assertion, we set $\tilde{\mathrm{P}}(y = 1|x) := 1/2$ for $x \in \tilde{X} \setminus X$. Then $\tilde{\mathrm{P}}(y = 1|x) \neq 1/2$ implies $x \in X$, and hence the classes $\tilde{X}_{-1}$ and $\tilde{X}_1$ of $\tilde{\mathrm{P}}$ are contained in X. The associated distances to the decision boundaries thus satisfy $\tilde{\Delta}(x) = \Delta(x)$ for all $x \in X$, and hence we finally find

$$\tilde{\mathrm{P}}_X\big(\{x \in \tilde{X} : \tilde{\Delta}(x) < t\}\big) = \mathrm{P}_X\big(\{x \in X : \Delta(x) < t\}\big)\,. \qquad \square$$

Let us now present some elementary examples of distributions having a strictly positive margin exponent.

Example 8.8 (Classes with strictly positive distance). Let X be a metric space, P be a distribution on $X \times Y$, and η be a version of the posterior probability for which the associated classes X_{-1} and X_1 have strictly positive distance, i.e., $d(X_{-1}, X_1) > 0$, and satisfy $\mathrm{P}_X(X_{-1} \cup X_1) = 1$. Then P has margin exponent α for all $\alpha > 0$.

To check this, let us write $t_0 := d(X_{-1}, X_1)$. Then we have $t_0 > 0$ and $\Delta(x) \geq t_0$ for all $x \in X_{-1} \cup X_1$. For $t \in (0, t_0]$, we thus find

$$\mathrm{P}_X\big(\{x \in X : \Delta(x) < t\}\big) = \mathrm{P}_X\big(\{x \in X_{-1} \cup X_1 : \Delta(x) < t\}\big) = 0\,.$$

Moreover, for $t > t_0$, we obviously have $\mathrm{P}_X(\{x \in X : \Delta(x) < t\}) \leq 1 < t_0^{-\alpha} t^{\alpha}$, and hence we obtain the assertion.

Finally, note that this example in particular includes distributions P on discrete metric spaces for which $\mathrm{P}(\{x \in X : \eta(x) = 1/2\}) = 0$. $\triangleleft$

Example 8.9 (Linear decision boundaries). Let $X \subset \mathbb{R}^d$ be a compact subset with strictly positive volume and P be a distribution on $X \times Y$ whose marginal distribution P_X is the uniform distribution. Moreover, assume that there exist a $w \in \mathbb{R}^d \setminus \{0\}$, a constant $b \in \mathbb{R}$, and a version η of the posterior probability such that the corresponding classes are given by $X_{-1} = \{x \in X : \langle w, x\rangle + b < 0\}$ and $X_1 = \{x \in X : \langle w, x\rangle + b > 0\}$. Then P has margin exponent $\alpha = 1$.

In order to check this, we first observe with the help of the rotation and translation invariance of the Lebesgue measure that we may assume without loss of generality that $w = e_1$ is the first vector of the standard ONB and $b = 0$. In addition, the compactness of X shows that there exists an $a > 0$ such that $X \subset [-a, a]^d$. Then we have

$$\begin{aligned}\{x \in X : \Delta(x) < t\} &= \{x \in X : -t < \langle e_1, x\rangle < t\} \\ &\subset \{x \in [-a,a]^d : -t < \langle e_1, x\rangle < t\},\end{aligned}$$

and since the volume of the last set is given by $a^{d-1}\min\{a,t\}$, the assertion becomes obvious.

Finally, note that P still has margin exponent $\alpha = 1$ if we assume that the classes X_{-1} and X_1 are "stripes", i.e., that they are described by finitely many parallel affine hyperplanes. ◁

Example 8.10 (Circular decision boundaries). Let $X \subset \mathbb{R}^d$ be a compact subset with a non-empty interior and P be a distribution on $X \times Y$ whose marginal distribution P_X is the uniform distribution. Moreover, assume that there exist an $x_0 \in \mathbb{R}^d$, an $r > 0$, and a version η of the posterior probability such that the corresponding classes are given by $X_{-1} = \{x \in X : \|x - x_0\| < r\}$ and $X_1 = \{x \in X : \|x - x_0\| > r\}$. Then P has margin exponent $\alpha = 1$.

To see this, we observe that we may assume without loss of generality that $x_0 = 0$. In addition, the compactness of X shows that there exists an $r_0 > 0$ such that $X \subset r_0 B_{\ell_2^d}$. Then we have

$$\begin{aligned}\{x \in X : \Delta(x) < t\} &= \{x \in X : r - t < \|x\|_2 < r + t\} \\ &\subset \{x \in r_0 B_{\ell_2^d} : r - t < \|x\|_2 < r + t\}.\end{aligned}$$

Since a simple calculation shows $(r+t)^d - (r-t)^d \le ct$ for a suitable constant $c > 0$ and all sufficiently small $t > 0$, we then obtain the assertion.

Finally note that P still has margin exponent $\alpha = 1$ if we assume that the classes X_{-1} and X_1 are described by finitely many circles. ◁

The previous two examples have not considered the *shape* of the input space X. However, this shape can have a substantial influence on the margin exponent. We refer to Exercise 8.3 for an example in this direction.

Let us finally consider an example of a distribution that does *not* have a strictly positive margin exponent.

Example 8.11 (Only trivial margin exponent). Assume that μ_{-1} is the uniform distribution on $[0,1]^2$ and that μ_1 is the uniform distribution on $\{0\} \times [0,1]$. Moreover, let $\eta := \mathbf{1}_{\{0\}\times[0,1]}$ and $P_X := (\mu_{-1} + \mu_1)/2$. Then the corresponding distribution P on $\mathbb{R}^2 \times \{-1,1\}$ has only margin exponent $\alpha = 0$. Indeed, for this version η of the posterior probability, we have $\Delta(x) = 0$ for all $x \in \{0\} \times [0,1]$, and it is easy to see that this cannot be changed by considering another version of the posterior probability. ◁

So far, we have only seen elementary examples of distributions having a non-trivial margin exponent. However, by combining these examples with the following lemma and the subsequent example, it is easy to see that the set of distributions having a non-trivial margin exponent is quite rich.

Lemma 8.12 (Images of inverse Hölder maps). *Let (X_1, d_1) and (X_2, d_2) be metric spaces and $\Psi : X_1 \to X_2$ be a measurable map whose image $\Psi(X_1)$ is measurable. Assume that Ψ is inverse Hölder continuous, i.e., there exist constants $c > 0$ and $\gamma \in (0, 1]$ such that*

$$d_1(x_1, x_1') \;\le\; c\, d_2^{\gamma}(\Psi(x_1), \Psi(x_1'))\,, \qquad x_1, x_1' \in X_1. \tag{8.4}$$

Let $Y := \{-1, 1\}$ and Q be a distribution on $X_1 \times Y$ that has some margin exponent $\alpha > 0$ for the version η_{Q} of its posterior probability. Furthermore, let P be a distribution on $X_2 \times Y$ for which there exists a constant $C \ge 1$ such that

$$\mathrm{P}_X(A) \;\le\; C\, \mathrm{Q}_X\bigl(\Psi^{-1}(A)\bigr) \tag{8.5}$$

for all measurable $A \subset X_2$. If P has a version η_{P} of its posterior probability such that

$$\eta_{\mathrm{P}}(\Psi(x_1)) = \eta_{\mathrm{Q}}(x_1)\,, \qquad x_1 \in X_1, \tag{8.6}$$

then P has margin exponent $\alpha\gamma$.

Proof. Without loss of generality, we may assume that $\eta_{\mathrm{P}}(x_2) = 1/2$ for all $x_2 \in X_2 \setminus \Psi(X_1)$. This assumption immediately implies $\Delta_{\mathrm{P}}(x_2) = 0$ for all $x_2 \in X_2 \setminus \Psi(X_1)$, but since (8.5) implies $\mathrm{P}_X(X_2 \setminus \Psi(X_1)) = 0$, we notice for later use that the behavior of Δ_{P} on $X_2 \setminus \Psi(X_1)$ has no influence on the margin exponent of P. Moreover, (8.4) implies that Ψ is injective and hence the inverse $\Psi^{-1} : \Psi(X_1) \to X_1$ of $\Psi : X_1 \to \Psi(X_1)$ exists. By (8.6), we conclude that $\eta_{\mathrm{P}}(x_2) = \eta_{\mathrm{Q}}(\Psi^{-1}(x_2)$ for all $x_2 \in \Psi(X_1)$. Let us now fix $x_2, x_2' \in X_2$ with $\eta_{\mathrm{P}}(x_2) < 1/2$ and $\eta_{\mathrm{P}}(x_2') > 1/2$. Then $\eta_{\mathrm{P}} \equiv 1/2$ on $X_2 \setminus \Psi(X_1)$ implies $x_2, x_2' \in \Psi(X_1)$, and hence we have $\eta_{\mathrm{Q}}(\Psi^{-1}(x_2)) < 1/2$ and $\eta_{\mathrm{Q}}(\Psi^{-1}(x_2')) > 1/2$. From this we conclude that

$$c\, d_2^{\gamma}(x_2, x_2') \ge d_1(\Psi^{-1}(x_2), \Psi^{-1}(x_2')) \ge \Delta_{\mathrm{Q}}(\Psi^{-1}(x_2))\,,$$

and hence $c\Delta_{\mathrm{P}}^{\gamma}(x_2) \ge \Delta_{\mathrm{Q}}(\Psi^{-1}(x_2))$. Repeating the argument above, we further see that this inequality hold not only for x_2 satisfying $\eta_{\mathrm{P}}(x_2) < 1/2$ but actually for all $x_2 \in \Psi(X_1)$. Combining this with our previous considerations and (8.5), we thus obtain

$$\begin{aligned}
\mathrm{P}_X\bigl(\{x_2 \in X_2 : \Delta_{\mathrm{P}}(x_2) < t\}\bigr) &= \mathrm{P}_X\bigl(\{x_2 \in \Psi(X_1) : \Delta_{\mathrm{P}}(x_2) < t\}\bigr) \\
&\le C\, \mathrm{Q}_X\bigl(\{x_1 \in X_1 : \Delta_{\mathrm{P}}(\Psi(x_1)) < t\}\bigr) \\
&\le C\, \mathrm{Q}_X\bigl(\{x_1 \in X_1 : \Delta_{\mathrm{Q}}(x_1) < ct^{\gamma}\}\bigr)\,.
\end{aligned}$$

Using the margin exponent of Q, we then find the assertion. □

Note that if the map Ψ in the lemma above is continuous and X_1 is a *complete* metric space, a simple Cauchy sequence argument combined with (8.4) shows that $\Psi(X_1)$ is a closed of subset X_2 and hence measurable.

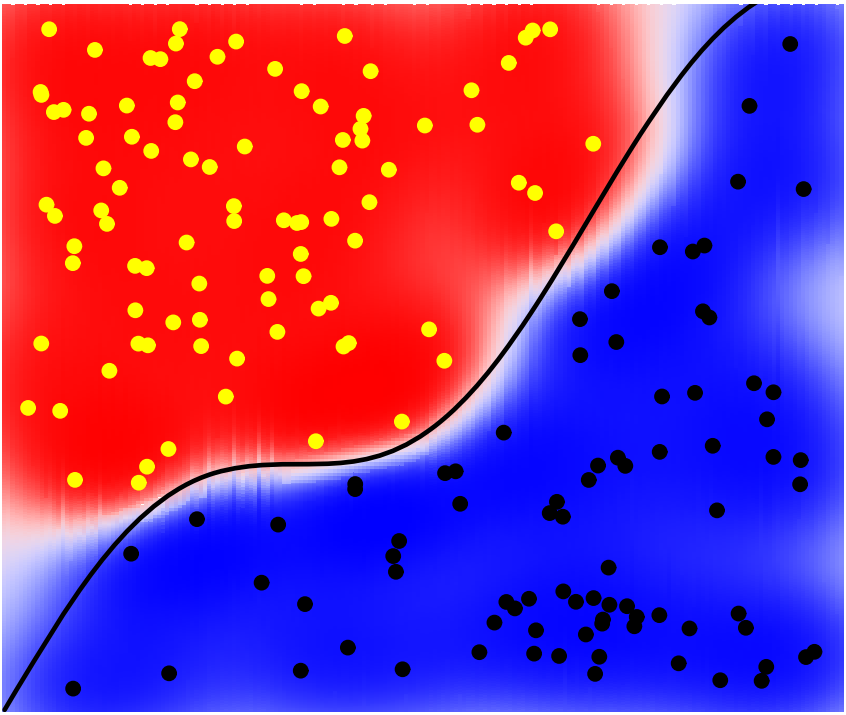

Fig. 8.1. Example of a distribution that has a large margin exponent. The red and blue areas are the negative and positive classes, respectively, and the decision boundary is located in the white area. The lighter colors indicate a low concentration of P_X, and hence only a few samples (indicated by yellow and black dots for positive and negative labels, respectively) can be found in this region. The black line shows an estimate of the decision boundary made by an SVM. Note that the larger the light area is, the larger the noise exponent is, and hence the better the SVM can estimate the true decision boundary. However, even in the absence of a light area, the corresponding distribution would have margin exponent $\alpha = 1$, if, for example, P_X had a bounded density.

If P_X is the image measure $\Psi(\mathrm{Q}_X)$ of Q_X under Ψ, then (8.5) becomes an equality with $C = 1$. However, this is by no means the only case where (8.5) can hold. For example, assume that P_X is only absolutely continuous with respect to $\Psi(\mathrm{Q}_X)$. Then (8.5) is obviously satisfied if the density of P_X with respect to $\Psi(\mathrm{Q}_X)$ is bounded. The next example illustrates such a situation.

Example 8.13 (Smooth transformations). Let $U, V \subset \mathbb{R}^d$ be two open non-empty sets and $\Psi : U \to V$ be a continuously differentiable map. Then recall that $\Psi(B)$ is measurable for all measurable $B \subset U$ and that a version of Sard's inequality (see, e.g., p. 313 in Floret, 1981) states that

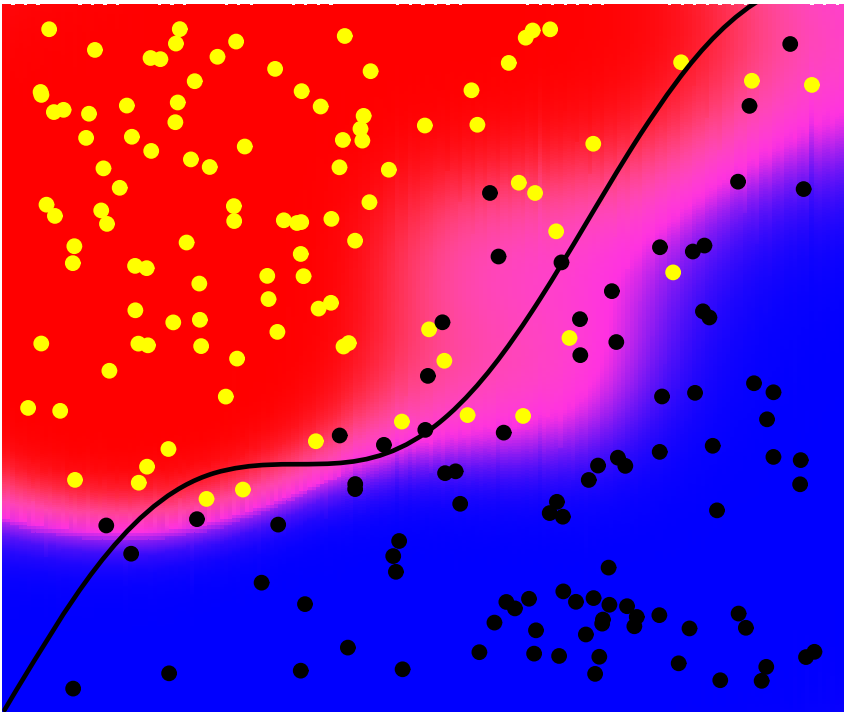

Fig. 8.2. Example of a distribution that has a large amount of noise around its decision boundary. The red and blue areas are the negative and positive classes, respectively, and the mixture of these colors indicates the amount of noise. Consequently, we see nearby samples with different labels (indicated by yellow and black dots for positive and negative labels, respectively) in this region. The black line shows an estimate of the decision boundary made by an SVM, whereas the true decision boundary is located in the middle of the purple region. Note that the larger the purple area around the decision boundary is, the larger the margin-noise exponent is. However, even there was no purple area the corresponding distribution would have margin-noise exponent $\alpha = 1$, if, for example, P_X had a bounded density.

$$\mathrm{vol}_d(\Psi(B)) \leq \int_B |\det \Psi'(x)| dx\,. \tag{8.7}$$

Let us further assume that Q is a distribution on $U \times Y$, where $Y := \{-1, 1\}$, such that $\operatorname{supp} \mathrm{Q}_X$ is bounded and $\mathrm{vol}_d(\Psi(\operatorname{supp} \mathrm{Q}_X)) > 0$. Then (8.7) yields $\mathrm{vol}_d(\operatorname{supp} \mathrm{Q}_X) > 0$. Let us assume for the sake of simplicity that Q_X is the uniform distribution on $\operatorname{supp} \mathrm{Q}_X$ and that P is a distribution on $V \times Y$ such that P_X is the uniform distribution on the compact set $\Psi(\operatorname{supp} \mathrm{Q}_X)$. Then (8.7) yields

$$\begin{aligned} \mathrm{P}_X(A) &\leq \frac{\mathrm{vol}_d(\operatorname{supp}\mathrm{Q}_X)}{\mathrm{vol}_d(\Psi(\operatorname{supp}\mathrm{Q}_X))} \int_U \mathbf{1}_A(\Psi(x))\,|\det \Psi'(x)|\mathrm{Q}_X(x) \\ &\leq \frac{\mathrm{vol}_d(\operatorname{supp}\mathrm{Q}_X)}{\mathrm{vol}_d(\Psi(\operatorname{supp}\mathrm{Q}_X))} \sup_{x\in\operatorname{supp}\mathrm{Q}_X} |\det \Psi'(x)| \cdot \mathrm{Q}_X(\Psi^{-1}(A)) \end{aligned}$$

for all measurable $A \subset V$. Since $\operatorname{supp}\mathrm{Q}_X$ is compact and $x \mapsto |\det \Psi'(x)|$ is continuous, we hence obtain (8.5).

Let us finally assume that Ψ is a diffeomorphism, i.e., Ψ is also bijective and $\Psi^{-1} : V \to U$ is continuously differentiable. If $V = \Psi(U)$ is convex, then the mean value theorem states that

$$\left\| \Psi^{-1}(x) - \Psi^{-1}(x') \right\|_2 \;\leq\; \sup_{t\in[0,1]} \left\| (\Psi^{-1})'(tx + (1-t)x') \right\|_2 \cdot \|x - x'\|_2$$

for all $x, x' \in V$. Since $\overline{\operatorname{co}\Psi(\operatorname{supp}\mathrm{Q}_X)} \subset V$ is compact, we thus obtain (8.4) for $\gamma := 1$, $X_1 := \operatorname{supp}\mathrm{Q}_X$, and $X_2 := \Psi(\operatorname{supp}\mathrm{Q}_X)$. Consequently, P has margin exponent α if Q has margin exponent α and (8.6) is satisfied. ◁

Our next goal is to show that small values of densities in the vicinity of the decision boundary improve the margin exponent. To this end, we say that the distributions P and Q on $X \times Y$ **generate the same classes** for the versions η_P and η_Q of their posterior probabilities of P and Q if $(2\eta_\mathrm{P} - 1)(2\eta_\mathrm{Q} - 1) > 0$. Obviously, in this case, the associated classes do coincide and hence the associated distances Δ_P and Δ_Q to the decision boundaries are equal. For such distributions, we can now prove the following lemma.

Lemma 8.14 (Low densities near the decision boundary). *Let (X, d) be a metric space and* Q *and* P *be two distributions on $X \times Y$ that generate the same classes for their posterior probabilities η_P and η_Q. We write Δ_Q for the associated distance to the decision boundary of* Q*. Furthermore, assume that P_X has a density $h : X \to [0,\infty)$ with respect to Q_X such that there exist constants $c > 0$ and $\gamma \in [0,\infty)$ satisfying*

$$h(x) \;\leq\; c\Delta_\mathrm{Q}^\gamma(x)\,, \qquad\qquad x \in X. \tag{8.8}$$

If Q *has margin exponent $\alpha \in [0,\infty)$ for Δ_Q, then* P *has margin exponent $\alpha + \gamma$.*

Proof. We have already seen before this lemma that there is a version Δ_P of the distance to the decision boundary of P that equals Δ_Q. For $t \geq 0$, we consequently obtain

$$\begin{aligned} \mathrm{P}_X\big(\{x \in X : \Delta_\mathrm{P}(x) < t\}\big) &= \int_{\Delta_\mathrm{Q}(x)<t} h(x)\, d\mathrm{Q}_X(x) \\ &\leq c\, t^\gamma\, \mathrm{Q}_X\big(\{x \in X : \Delta_\mathrm{Q}(x) < t\}\big)\,. \end{aligned}$$

From this we immediately obtain the assertion. □

For classifying with the hinge loss, we will see below that it is more suitable to measure the size of $\{x \in X : \Delta(x) < t\}$ by $|2\eta - 1|\mathrm{P}_X$ instead of P_X. This motivates the following definition.

Definition 8.15. *Let (X, d) be a metric space and P be a distribution on $X \times Y$. We say that P has* ***margin-noise exponent*** *$\beta \in [0, \infty)$ for the version $\eta : X \to [0, 1]$ of its posterior probability if there exists a constant $c \geq 1$ such that the distance to the decision boundary Δ associated to η satisfies*

$$\int_{\Delta(x)<t} |2\eta(x) - 1| \, d\mathrm{P}_X(x) \;\leq\; c\, t^{\beta}\,, \qquad\qquad t \geq 0. \tag{8.9}$$

Let us now investigate the relation between the margin exponent and the margin-noise exponent. Lemma 8.14 suggests that to this end we need to describe how the distance to the decision boundary influences the amount of noise. This is done in the following definition.

Definition 8.16. *Let (X, d) be a metric space, P be a distribution on $X \times Y$, and $\eta : X \to [0, 1]$ be a version of its posterior probability. We say that the associated* ***distance to the decision boundary Δ controls the noise by the exponent*** *$\gamma \in [0, \infty)$ if there exists a constant $c > 0$ such that*

$$|2\eta(x) - 1| \;\leq\; c\Delta^{\gamma}(x) \tag{8.10}$$

for P_X-almost all $x \in X$.

Note that since $|2\eta(x) - 1| \leq 1$ for all $x \in X$, condition (8.10) becomes trivial whenever $\Delta(x) \geq c^{-1/\gamma}$. Consequently, (8.10) is a condition that only considers points $x \in X$ with sufficiently small distance to the opposite class. In simple words, it states that $\eta(x)$ is close to the level $1/2$ of "complete noise" if x approaches the decision boundary. The following lemma, whose omitted proof is almost identical to that of Lemma 8.14, now relates the margin exponent to the margin-noise exponent.

Lemma 8.17. *Let X be a metric space and P be a distribution on $X \times Y$ that has margin exponent $\alpha \in [0, \infty)$ for the version η of its posterior probability. Assume that the associated distance to the decision boundary Δ controls the noise by the exponent $\gamma \in [0, \infty)$. Then P has margin-noise exponent $\alpha + \gamma$.*

Note that for $\alpha = 0$ the preceding lemma states that every distribution whose distance to the decision boundary controls the noise by some exponent $\gamma > 0$ has margin-noise exponent γ. In other words, distributions that have a high amount of noise around their decision boundary have a non-trivial margin-noise exponent. We refer to Figure 8.2 for an illustration of this situation. In addition, note that in the case $\gamma = 0$ the lemma states that distributions having some margin exponent $\alpha > 0$ also have margin-noise exponent α. Consequently, all considerations on the margin exponent made so

far have an immediate consequence for the margin-noise exponent. Finally, the general message of Lemma 8.17 is that for distributions satisfying both assumptions their exponents add up to become the margin-noise exponent.

With the help of the concepts above, we can now return to our initial goal of estimating the approximation error function for Gaussian kernels.

Theorem 8.18 (Approximation error of Gaussian kernels). *Let L be the hinge loss and P be a distribution on $\mathbb{R}^d \times \{-1,1\}$ that has margin-noise exponent $\beta \in (0,\infty)$ and whose marginal distribution P_X has tail exponent $\tau \in (0,\infty]$. Then there exist constants $c_{d,\tau} > 0$ and $\tilde{c}_{d,\beta} > 0$ such that for all $\gamma > 0$ and all $\lambda > 0$ there exists a function $f^* \in H_\gamma(\mathbb{R}^d)$ in the RKHS $H_\gamma(\mathbb{R}^d)$ of the Gaussian RBF kernel k_γ such that $\|f^*\|_\infty \le 1$ and*

$$\lambda \|f^*\|^2_{H_\gamma(\mathbb{R}^d)} + \mathcal{R}_{L,\mathrm{P}}(f^*) - \mathcal{R}^*_{L,\mathrm{P}} \le c_{d,\tau}\, \lambda^{\frac{\tau}{d+\tau}}\, \gamma^{-\frac{d\tau}{d+\tau}} \,+\, \tilde{c}_{d,\beta}\, c\, \gamma^\beta\,,$$

where c is the constant appearing in (8.9). In particular, we have

$$A_2^{(\gamma)}(\lambda) \;\le\; \max\bigl\{c_{d,\tau}, \tilde{c}_{d,\beta}\, c\bigr\} \cdot \bigl(\lambda^{\frac{\tau}{d+\tau}} \gamma^{-\frac{d\tau}{d+\tau}} + \gamma^\beta\bigr)\,.$$

Proof. Obviously, it suffices to prove the first assertion. Let $\eta : \mathbb{R}^d \to [0,1]$ be a version of the posterior probability of P such that its associated $\Delta(x)$, $x \in \mathbb{R}^d$, satisfies (8.9). We define X_{-1} and X_1 as in Definition 8.5 and further use the shorthand $B := \{x \in \mathbb{R}^d : \|x\|_2 \le 1\}$ for the closed unit ball of $\mathbb{R}^d$. Let us fix a $\rho > 0$ such that $X_{-1} \cap \rho B \ne \emptyset$ and $X_1 \cap \rho B \ne \emptyset$. Then an easy consideration shows that $\Delta(x) \le 2\rho$ for all $x \in \rho B$. Finally, we define $f_\rho : \mathbb{R}^d \to [-1,1]$ by

$$f_\rho(x) := \mathbf{1}_{(X_{-1} \cup X_1) \cap 3\rho B}(x) \cdot \operatorname{sign}\bigl(2\eta(x) - 1\bigr)\,, \qquad\qquad x \in \mathbb{R}^d,$$

and $g_\rho : \mathbb{R}^d \to \mathbb{R}$ by $g_\rho := (\pi\gamma^2)^{-d/4} f_\rho$. Obviously, we have $g_\rho \in L_2(\mathbb{R}^d)$ with

$$\|g_\rho\|_{L_2(\mathbb{R}^d)} \le \left(\frac{9\rho^2}{\pi\gamma^2}\right)^{\frac{d}{4}} \sqrt{\mathrm{vol}_d(B)}\,, \tag{8.11}$$

where $\mathrm{vol}_d(B)$ denotes the volume of B. Let us now recall Lemma 4.45 and (4.43), which showed that $L_2(\mathbb{R}^d)$ is a feature space of $H_\gamma(\mathbb{R}^d)$ with canonical metric surjection $V_\gamma : L_2(\mathbb{R}^d) \to H_\gamma(\mathbb{R}^d)$ given by

$$V_\gamma g(x) = \left(\frac{4}{\pi\gamma^2}\right)^{\frac{d}{4}} \int_{\mathbb{R}^d} e^{-2\gamma^{-2}\|x-x'\|_2^2}\, g(x')dx'\,, \qquad\qquad g \in L_2(\mathbb{R}^d),\, x \in \mathbb{R}^d.$$

By Theorem 4.21 and (8.11), we then obtain

$$\|V_\gamma g_\rho\|_{H_\gamma(\mathbb{R}^d)} \le \left(\frac{9\rho^2}{\pi\gamma^2}\right)^{\frac{d}{4}} \sqrt{\mathrm{vol}_d(B)}\,. \tag{8.12}$$

Moreover, $g_\rho = (\pi\gamma^2)^{-d/4} f_\rho$ together with $\|f_\rho\|_\infty \le 1$ and (A.3) implies

$$\left|V_\gamma g_\rho(x)\right| \le \left(\frac{2}{\pi\gamma^2}\right)^{\frac{d}{2}} \int_{\mathbb{R}^d} e^{-2\gamma^{-2}\|x-x'\|_2^2} dx' = 1$$

for all $x \in \mathbb{R}^d$. Therefore, Theorem 2.31 yields

$$\mathcal{R}_{L_{\text{hinge}},\mathrm{P}}(V_\gamma g_\rho) - \mathcal{R}^*_{L_{\text{hinge}},\mathrm{P}} \;=\; \mathbb{E}_{\mathrm{P}_X}\big(\,|V_\gamma g_\rho - f^*_{L_{\text{class}},\mathrm{P}}| \cdot |2\eta - 1|\,\big). \quad (8.13)$$

In order to bound $|V_\gamma g_\rho - f^*_{L_{\text{class}},\mathrm{P}}|$ we fix an $x \in X_1 \cap \rho B$. Furthermore, we write $B(x,r) := \{x' \in \mathbb{R}^d : \|x-x'\|_2 < r\}$ for the *open* ball with radius r and center x. For $x' \in B(x, \Delta(x))$, we then have $\|x - x'\|_2 < \Delta(x)$, and thus we obtain $x' \in X_1$. Moreover, we also have $\|x'\|_2 \le \|x-x'\|_2 + \|x\|_2 < \Delta(x) + \rho \le 3\rho$, and therefore we conclude that

$$B(x, \Delta(x)) \subset X_1 \cap 3\rho B\,.$$

Similarly, for $x' \in X_{-1}$, we have $\Delta(x) \le \|x - x'\|_2$, and hence we obtain

$$X_{-1} \cap 3\rho B \;\subset\; X_{-1} \;\subset\; \mathbb{R}^d \backslash B(x, \Delta(x))\,.$$

Combining the definition of f with these two inclusions and (A.3) now yields

$$\begin{aligned}
V_\gamma g_\rho(x) &= \left(\frac{2}{\pi\gamma^2}\right)^{\frac{d}{2}} \int_{\mathbb{R}^d} e^{-2\gamma^{-2}\|x-x'\|_2^2} f_\rho(x')dx' \\
&= \left(\frac{2}{\pi\gamma^2}\right)^{\frac{d}{2}} \left(\int_{X_1 \cap 3\rho B} e^{-2\gamma^{-2}\|x-x'\|_2^2} dx' - \int_{X_{-1}\cap 3\rho B} e^{-2\gamma^{-2}\|x-x'\|_2^2} dx' \right) \\
&\ge \left(\frac{2}{\pi\gamma^2}\right)^{\frac{d}{2}} \left(\int_{B(x,\Delta(x))} e^{-2\gamma^{-2}\|x-x'\|_2^2} dx' - \int_{\mathbb{R}^d\backslash B(x,\Delta(x))} e^{-2\gamma^{-2}\|x-x'\|_2^2} dx' \right) \\
&= 2\left(\frac{2}{\pi\gamma^2}\right)^{\frac{d}{2}} \int_{B(x,\Delta(x))} e^{-2\gamma^{-2}\|x-x'\|_2^2} dx' - 1\,.
\end{aligned}$$

Since $V_\gamma g_\rho(x) \le 1$ and $f^*_{L_{\text{class}},\mathrm{P}}(x) = 1$, we thus obtain

$$\begin{aligned}
|V_\gamma g_\rho(x) - f^*_{L_{\text{class}},\mathrm{P}}(x)| &= 1 - V_\gamma g_\rho(x) \\
&\le 2 - 2\left(\frac{2}{\pi\gamma^2}\right)^{\frac{d}{2}} \int_{B(x,\Delta(x))} e^{-2\gamma^{-2}\|x-x'\|_2^2} dx' \\
&= 2 - 2\left(\frac{2}{\pi\gamma^2}\right)^{\frac{d}{2}} \int_{B(0,\Delta(x))} e^{-2\gamma^{-2}\|x'\|_2^2} dx'\,.
\end{aligned}$$

Using the rotation invariance of $x' \mapsto e^{-2\gamma^{-2}\|x'\|_2^2}$ and $\Gamma(1+t) = t\Gamma(t)$, $t > 0$, we further find

$$\left(\frac{2}{\pi\gamma^2}\right)^{\frac{d}{2}} \int\limits_{B(0,\Delta(x))} e^{-2\gamma^{-2}\|x'\|_2^2} dx' = \frac{d}{\Gamma(1+d/2)} \left(\frac{2}{\gamma^2}\right)^{\frac{d}{2}} \int_0^{\Delta(x)} e^{-2\gamma^{-2}r^2} r^{d-1} dr$$

$$= \frac{2}{\Gamma(d/2)} \int_0^{\sqrt{2}\Delta(x)\gamma^{-1}} e^{-r^2} r^{d-1} dr$$

$$= \frac{1}{\Gamma(d/2)} \int_0^{2\Delta^2(x)\gamma^{-2}} e^{-r} r^{d/2-1} dr\,.$$

Combining this equation with the previous estimate yields

$$|V_\gamma g_\rho(x) - f^*_{L_{\text{class}},\mathrm{P}}(x)| \le 2 - \frac{2}{\Gamma(d/2)} \int_0^{2\Delta^2(x)\gamma^{-2}} e^{-r} r^{d/2-1} dr$$

$$= \frac{2}{\Gamma(d/2)} \left(\int_0^\infty e^{-r} r^{d/2-1} dr - \int_0^{2\Delta^2(x)\gamma^{-2}} e^{-r} r^{d/2-1} dr \right)$$

$$= \frac{2}{\Gamma(d/2)} \int_0^\infty \mathbf{1}_{(2\Delta^2(x)\gamma^{-2},\infty)}(r)\, e^{-r} r^{d/2-1} dr$$

for all $x \in X_1 \cap \rho B$. Moreover, repeating the proof above for $x \in X_{-1} \cap \rho B$, we see that this estimate actually holds for all $x \in (X_{-1} \cup X_1) \cap \rho B$. Since $|2\eta(x) - 1| = 0$ for $x \notin X_{-1} \cup X_1$, we thus find

$$\begin{aligned}
&\int_{\rho B} |V_\gamma g_\rho(x) - f^*_{L_{\text{class}},\mathrm{P}}(x)| \cdot |2\eta(x) - 1|\, d\mathrm{P}_X(x) \\
&\le \frac{2}{\Gamma(d/2)} \int_{X_{-1}\cup X_1} \int_0^\infty \mathbf{1}_{(2\Delta^2(x)\gamma^{-2},\infty)}(r)\, e^{-r} r^{d/2-1} |2\eta(x) - 1|\, dr\, d\mathrm{P}_X(x) \\
&= \frac{2}{\Gamma(d/2)} \int_0^\infty e^{-r} r^{d/2-1} \int_{\mathbb{R}^d} \mathbf{1}_{[0,\gamma(r/2)^{1/2})}(\Delta(x)) \cdot |2\eta(x) - 1|\, d\mathrm{P}_X(x)\, dr \\
&\le \frac{2^{1-\beta/2}\, c\, \gamma^\beta}{\Gamma(d/2)} \int_0^\infty e^{-r} r^{(\beta+d)/2-1} dr \\
&= \frac{2^{1-\beta/2}\, c\, \Gamma((\beta+d)/2)}{\Gamma(d/2)}\, \gamma^\beta
\end{aligned}$$

by the definition of the margin-noise exponent. Using equation (8.13) together with $\|V_\gamma g_\rho\|_\infty \le 1$ and the tail exponent inequality (7.61) thus yields

$$\begin{aligned}
&\mathcal{R}_{L_{\text{hinge}},\mathrm{P}}(V_\gamma g_\rho) - \mathcal{R}^*_{L_{\text{hinge}},\mathrm{P}} \\
&\le \int_{\rho B} |V_\gamma g_\rho(x) - f^*_{L_{\text{class}},\mathrm{P}}(x)| \cdot |2\eta(x) - 1|\, d\mathrm{P}_X(x) + 2\mathrm{P}_X(\mathbb{R}^d \setminus \rho B) \\
&\le \frac{2^{1-\beta/2}\, c\, \Gamma((\beta+d)/2)}{\Gamma(d/2)}\, \gamma^\beta + 2\rho^{-\tau}\,. \qquad (8.14)
\end{aligned}$$

By combining this estimate with (8.12), we therefore obtain

$$\begin{aligned}
&\lambda \|V_\gamma g_\rho\|_{H_\gamma}^2 + \mathcal{R}_{L_{\text{hinge}},\mathrm{P}}(V_\gamma g_\rho) - \mathcal{R}^*_{L_{\text{hinge}},\mathrm{P}} \\
&\leq \lambda \Big(\frac{9\rho^2}{\pi\gamma^2}\Big)^{\frac{d}{2}} \operatorname{vol}_d(B) + \frac{2^{1-\beta/2}\, c\, \Gamma((\beta+d)/2)}{\Gamma(d/2)}\, \gamma^\beta + 2\rho^{-\tau}\,. \qquad (8.15)
\end{aligned}$$

So far, we have found a g_ρ satisfying (8.15) only if ρ satisfies both $X_{-1} \cap \rho B \neq \emptyset$ and $X_1 \cap \rho B \neq \emptyset$. Our next goal is to find such a g_ρ for the remaining $\rho > 0$. To this end, let us first consider a $\rho > 0$ such that $X_{-1} \cap \rho B = \emptyset$ and $X_1 \cap \rho B = \emptyset$. For such ρ, we set $f_\rho := g_\rho := 0$, so that (8.12) and (8.13) are trivially satisfied. Moreover, for $x \in \rho B$, our assumption on ρ guarantees $x \notin X_{-1} \cup X_1$, which in turn yields $|2\eta(x) - 1| = 0$. Consequently, we find

$$\begin{aligned}
&\mathcal{R}_{L_{\text{hinge}},\mathrm{P}}(V_\gamma g_\rho) - \mathcal{R}^*_{L_{\text{hinge}},\mathrm{P}} \\
&\leq \int_{\rho B} |V_\gamma g_\rho(x) - f^*_{L_{\text{class}},\mathrm{P}}(x)| \cdot |2\eta(x) - 1|\, d\mathrm{P}_X(x) + 2\mathrm{P}_X(\mathbb{R}^d \backslash \rho B) \\
&\leq 2\rho^{-\tau}\,,
\end{aligned}$$

and hence (8.15) turns out to be true for the ρ and g_ρ considered. Let us now consider a $\rho \geq 1$ such that $X_{-1} \cap \rho B = \emptyset$ and $X_1 \cap \rho B \neq \emptyset$. In this case, we define $f_\rho := \mathbf{1}_{2\rho B}$ and $g_\rho := (\pi\gamma^2)^{-d/4} f_\rho$. Then (8.12) and (8.13) are obviously satisfied. Moreover, for $x \in \rho B$, we have $B(x,\rho) \subset 2\rho B$, and hence repeating our previous calculations yields

$$\begin{aligned}
V_\gamma g_\rho(x) = \left(\frac{2}{\pi\gamma^2}\right)^{\frac{d}{2}} \int_{2\rho B} e^{-2\gamma^{-2}\|x-y\|_2^2} dy &\geq \left(\frac{2}{\pi\gamma^2}\right)^{\frac{d}{2}} \int_{B(0,\rho)} e^{-2\gamma^{-2}\|y\|_2^2} dy \\
&= \frac{1}{\Gamma(d/2)} \int_0^{2\rho^2\gamma^{-2}} e^{-r} r^{d/2-1} dr\,.
\end{aligned}$$

For $x \in X_1 \cap \rho B$, it is then easy to conclude that

$$\begin{aligned}
|V_\gamma g_\rho(x) - f^*_{L_{\text{class}},\mathrm{P}}(x)| &\leq 1 - \frac{1}{\Gamma(d/2)} \int_0^{2\rho^2\gamma^{-2}} e^{-r} r^{d/2-1} dr \\
&= \frac{1}{\Gamma(d/2)} \int_{2\rho^2\gamma^{-2}}^{\infty} e^{-r} r^{d/2-1} dr \\
&\leq \frac{2^{-\beta/2}\, \Gamma((\beta+d)/2)}{\Gamma(d/2)}\, \gamma^\beta\,,
\end{aligned}$$

where in the last step we used the last estimate of Lemma A.1.1. Since the constant c in (8.9) is assumed to be not smaller than 1, we then conclude that (8.15) holds. Finally, if $\rho \in (0,1)$, the latter inequality is satisfied for $g_\rho := 0$, and consequently we have found for all $\rho > 0$ a function g_ρ such that (8.15) holds. Minimizing (8.15) with respect to ρ now yields the assertion. □

Having a bound on the approximation error, we can now derive learning rates for λ_n and γ_n chosen *a priori* with the help of Theorem 8.1 and Theorem 8.3. However, it turns out that the optimal rates of such an approach

require knowledge about the distribution, namely the margin-noise exponent and the tail exponent. Since in practice such knowledge is not available, we skip the derivation of these rates and focus directly on data-dependent adaptive parameter selection strategies. The strategy we will focus on is a simple modification of the TV-SVM considered in Section 6.5. This modification determines not only the regularization parameter λ by a grid but also the kernel parameter γ. As in Section 6.5, we therefore need to know how much the approximation error function is affected by small changes of λ and γ. This is investigated in the following corollary.

Corollary 8.19. *Let* P *be a distribution on* $\mathbb{R}^d \times \{-1,1\}$ *that has margin-noise exponent* $\beta \in (0,\infty)$ *and whose* P_X *has tail exponent* $\tau \in (0,\infty]$*. Moreover, for* $\varepsilon > 0$ *and* $\delta > 0$*, we fix a finite* ε*-net* $\Lambda \subset (0,1]$ *and a finite* δ*-net* $\Gamma \subset (0,1]$*, respectively. Then, for all* $p > 0$*,* $q \geq 0$*, and all* $x \in (0,1]$*, we have*

$$\min_{\lambda\in\Lambda,\,\gamma\in\Gamma} \left(A_2^{(\gamma)}(\lambda) + x\lambda^{-p}\gamma^{-q} \right) \leq c \left(x^{\frac{\beta\tau}{\beta\tau+d\beta p+\beta p\tau+dp\tau+q\tau}} + \varepsilon^{\frac{\tau}{d+\tau}} + \delta^{\beta} \right),$$

where $c \geq 1$ *is a constant independent of* x*,* Λ*,* ε*,* Γ*, and* δ*.*

The proof of the preceding corollary actually yields a particular expression for the constant c, but since this expression is extremely complicated, we decided to omit the details.

Proof. Without loss of generality, we may assume that Λ and Γ are of the form $\Lambda = \{\lambda_1, \ldots, \lambda_m\}$ and $\Gamma = \{\gamma_1, \ldots, \gamma_\ell\}$ with $\lambda_{i-1} < \lambda_i$ and $\gamma_{j-1} < \gamma_j$ for all $i = 2, \ldots, m$ and $\gamma = 2, \ldots, \ell$, respectively. Moreover, we fix a minimizer (λ^*, γ^*) of the function $(\lambda, \gamma) \mapsto \lambda^{\frac{\tau}{d+\tau}} \gamma^{-\frac{d\tau}{d+\tau}} + \gamma^\beta + x\lambda^{-p}\gamma^{-q}$ defined on $[0,1]^2$. Analogously to the proof of Lemma 6.30, we then see that there exist indexes $i \in \{1, \ldots, m\}$ and $j \in \{1, \ldots, \ell\}$ such that $\lambda^* \leq \lambda_i \leq \lambda^* + 2\varepsilon$ and $\gamma^* \leq \gamma_j \leq \gamma^* + 2\delta$. By Theorem 8.18, we then obtain

$$\begin{aligned}
&\min_{\lambda\in\Lambda,\,\gamma\in\Gamma} \left(A_2^{(\gamma)}(\lambda) + x\lambda^{-p}\gamma^{-q} \right) \\
&\leq A_2^{(\gamma_j)}(\lambda_i) + x\lambda_i^{-p}\gamma_j^{-q} \\
&\leq c_1 \left(\lambda_i^{\frac{\tau}{d+\tau}} \gamma_j^{-\frac{d\tau}{d+\tau}} + \gamma_j^{\beta} \right) + x\,(\lambda^*)^{-p}(\gamma^*)^{-q} \\
&\leq c_1 \left((\lambda^* + 2\varepsilon)^{\frac{\tau}{d+\tau}} (\gamma^*)^{-\frac{d\tau}{d+\tau}} + (\gamma^* + 2\delta)^{\beta} \right) + x\,(\lambda^*)^{-p}(\gamma^*)^{-q} \\
&\leq c_2 \left((\lambda^*)^{\frac{\tau}{d+\tau}} (\gamma^*)^{-\frac{d\tau}{d+\tau}} + (\gamma^*)^{\beta} + x\,(\lambda^*)^{-p}(\gamma^*)^{-q} + \varepsilon^{\frac{\tau}{d+\tau}} + \delta^{\beta} \right) \\
&= c_2 \min_{\lambda,\gamma\in[0,1]} \left(\lambda^{\frac{\tau}{d+\tau}} \gamma^{-\frac{d\tau}{d+\tau}} + \gamma^{\beta} + x\lambda^{-p}\gamma^{-q} \right) + c_2\varepsilon^{\frac{\tau}{d+\tau}} + c_2\delta^{\beta},
\end{aligned}$$

where in the second to last step we used $\gamma^* \leq 1$, and where c_1 and c_2 are suitable constants independent of x, Λ, ε, Γ, and δ. Now the assertion follows from Lemma A.1.6. □

Let us now introduce the announced modification of the TV-SVM.

Definition 8.20. *Let $\Lambda := (\Lambda_n)$ and $\Gamma := (\Gamma_n)$ be sequences of finite subsets $\Lambda_n, \Gamma_n \subset (0,1]$. For $D := ((x_1, y_1), \ldots, (x_n, y_n)) \in (\mathbb{R}^d \times \{-1,1\})^n$, we define*

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

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

$$f_{\mathrm{D}_1,\lambda,\gamma} := \arg\min_{f \in H_\gamma} \lambda \|f\|_{H_\gamma}^2 + \mathcal{R}_{L_{\mathrm{hinge}},\mathrm{D}_1}(f), \qquad (\lambda,\gamma) \in \Lambda_n \times \Gamma_n,$$

and use D_2 to determine (λ,γ) by choosing a $(\lambda_{D_2}, \gamma_{D_2}) \in \Lambda_n \times \Gamma_n$ such that

$$\mathcal{R}_{L_{\mathrm{class}},\mathrm{D}_2}(f_{\mathrm{D}_1,\lambda_{D_2},\gamma_{D_2}}) = \min_{(\lambda,\gamma)\in\Lambda_n\times\Gamma_n} \mathcal{R}_{L_{\mathrm{class}},\mathrm{D}_2}(f_{\mathrm{D}_1,\lambda,\gamma}).$$

A learning method that produces such decision functions $f_{\mathrm{D}_1,\lambda_{D_2},\gamma_{D_2}}$ for all $D \in (X \times Y)^n$ is called a ***training validation support vector machine (TV-SVM)*** *with respect to Λ and Γ.*

Note that in the parameter selection step we consider the classification loss, although in principle we could have also used the hinge loss. Since the classification risk only considers the sign of a decision function and not its values, it is not necessary to clip $f_{\mathrm{D}_1,\lambda,\gamma}$ in this parameter selection step. Finally note that in general the pair $(\lambda_{D_2}, \gamma_{D_2})$, and thus also $f_{\mathrm{D}_1,\lambda_{D_2},\gamma_{D_2}}$, is not uniquely determined.

The existence of a measurable TV-SVM with respect to Λ and Γ can be shown by elementary modifications of the proof of Lemma 6.29, which established the measurability of TV-SVMs with respect to Λ. We omit the details and leave the proof to the interested reader. Moreover, the oracle inequalities established in Theorems 6.31 and 6.32 can easily be adapted to the new TV-SVM, too. Again, we skip the details and only present the resulting consistency and learning rates.

Theorem 8.21. *Let $\Lambda := (\Lambda_n)$ and $\Gamma := (\Gamma_n)$ be sequences of finite subsets $\Lambda_n, \Gamma_n \subset (0,1]$ such that Λ_n is an $n^{-1/2}$-net of $(0,1]$ and Γ_n is an $n^{-1/(2d)}$-net of $(0,1]$, respectively. Furthermore, assume that the cardinalities $|\Lambda_n|$ and $|\Gamma_n|$ grow polynomially in n. Then every measurable TV-SVM with respect to Λ and Γ is universally consistent. Moreover, for distributions on $\mathbb{R}^d \times \{-1,1\}$ that have margin-noise exponent $\beta \in (0,\infty)$ and whose P_X have tail exponent $\tau \in (0,\infty]$, such TV-SVMs learn with rate $n^{-\gamma}$ where*

$$\gamma := \begin{cases} \frac{\beta\tau}{4\beta\tau + 2d\beta + 2d\tau} & \text{if } \tau < \infty \\ \frac{\beta}{3\beta + 2d} + \rho & \text{if } \tau = \infty \end{cases}$$

and $\rho > 0$ is an arbitrarily small number.

Proof. Since the proof closely follows the lines of the proofs of Theorem 6.31 and 6.32, we only mention the main steps. To show the consistency, we need a simple modification of Theorem 6.32 to address the second set of parameter candidates Γ_n and the fact that $\lim_{\lambda \to 0} A_2^{(\gamma)}(\lambda) = 0$ for any fixed γ. Moreover, the first rate follows from combining such a modification of Theorem 6.32 with Corollary 8.19. For the last rate, we first establish an oracle inequality for the TV-SVM using Theorem 8.3 and a simple adaptation of the proof of Theorem 6.31. This oracle inequality is then combined with Corollary 8.19. □

8.3 Advanced Concentration Results for SVMs (*)

In this section, we improve the learning rates obtained in the previous section with the help of the advanced concentration results of Chapter 7. To this end, let us first recall that one of the key ingredients of that chapter was a *variance bound*. Consequently, our first goal is to establish such a bound for the hinge loss. We begin with the following definition.

Definition 8.22. *A distribution* P *on* $X \times \{-1, 1\}$ *is said to have* ***noise exponent*** $q \in [0, \infty]$ *if there exists a constant* $c > 0$ *such that*

$$\mathrm{P}_X\big(\{x \in X : |2\eta(x) - 1| < t\}\big) \;\leq\; (c\,t)^q\,, \qquad\qquad t \geq 0. \tag{8.16}$$

Note that we have a high amount of noise in the labeling process at $x \in X$ if $\eta(x)$ is close to $1/2$, i.e., if $|2\eta(x) - 1|$ is close to 0. Consequently, the noise exponent measures the size of the set of points that have a high noise in the labeling process. Obviously, every distribution has noise exponent $q = 0$, whereas noise exponent $q = \infty$ means that η is bounded away from the critical level $1/2$. In particular, noise-free distributions, i.e., distributions P with $\mathcal{R}^*_{L_{\mathrm{class}},\mathrm{P}} = 0$, have noise exponent $q = \infty$. Moreover, if P has noise exponent q, then P also has noise exponent q' for all $q' < q$. In addition, note that the noise exponent does *not* locate the points x having high noise, i.e., it does not consider their closeness to the decision boundary. Finally, it is important to note that, unlike the concepts we considered in Section 8.2, the noise exponent does *not* depend on a specific version of the posterior probability. This makes it technically easier to combine the noise exponent with the previously introduced concepts such as the margin-noise exponent. The next lemma, which relates the noise exponent to the margin-noise exponent, illustrates this.

Lemma 8.23 (Relation between noise exponents). *Let* (X, d) *be a metric space and* P *be a distribution on* $X \times Y$ *that has noise exponent* $q \in [0, \infty)$*. Furthermore, assume that there exists a version* η *of its posterior probability such that the associated distance to the decision boundary* Δ *controls the noise by the exponent* $\gamma \in [0, \infty)$ *in the sense of Definition 8.16. Then* P *has margin exponent* $\alpha := \gamma q$ *and margin-noise exponent* $\beta := \gamma(q + 1)$*.*

Proof. We have $\{x \in X : \Delta(x) < t\} \subset \{x \in X : |2\eta(x) - 1| < ct^\gamma\}$, where the inclusion is only P_X-almost surely and $c > 0$ is the constant appearing in (8.10). From this it is easy to conclude that P has margin exponent γq. Combining this with Lemma 8.17 yields the second assertion. □

Theorem 8.24 (Variance bound for the hinge loss). *Let* P *be a distribution on* $X \times Y$ *that has noise exponent* $q \in [0, \infty]$. *Moreover, let* $f^*_{L,\mathrm{P}} : X \to [-1, 1]$ *be a fixed Bayes decision function for the hinge loss* L. *Then, for all measurable* $f : X \to \mathbb{R}$, *we have*

$$\mathbb{E}_\mathrm{P}(L \circ \widehat{f} - L \circ f^*_{L,\mathrm{P}})^2 \;\le\; 6\, c^{q/(q+1)} \left(\mathbb{E}_\mathrm{P}(L \circ \widehat{f} - L \circ f^*_{L,\mathrm{P}})\right)^{q/(q+1)},$$

where c *is the constant appearing in (8.16).*

Proof. Since the range of $\widehat{f}$ is contained in $[-1, 1]$, we may restrict our considerations to functions $f : X \to [-1, 1]$. Now observe that for such f we have $L(y, f(x)) = 1 - yf(x)$ for all $x \in X$, $y = \pm 1$, and hence we obtain

$$\big(L(y, f(x)) - L(y, f^*_{L,\mathrm{P}}(x))\big)^2 = \big(yf^*_{L,\mathrm{P}}(x) - yf(x)\big)^2 = \big(f(x) - f^*_{L,\mathrm{P}}(x)\big)^2$$

for all $x \in X$, $y = \pm 1$. From this we conclude that

$$\begin{aligned}
\mathbb{E}_\mathrm{P}(L \circ f - L \circ f^*_{L,\mathrm{P}})^2 &= \int_X |f - f^*_{L,\mathrm{P}}|^2 \, d\mathrm{P}_X \\
&\le 2\int_{|2\eta-1|\ge s} |f - f^*_{L,\mathrm{P}}| \, d\mathrm{P}_X + 2\int_{|2\eta-1|<s} |f - f^*_{L,\mathrm{P}}| \, d\mathrm{P}_X \\
&\le 2s^{-1}\int_X |f - f^*_{L,\mathrm{P}}| \cdot |2\eta - 1| d\mathrm{P}_X + 4\mathrm{P}_X(|2\eta - 1| < s) \\
&\le 2s^{-1}\mathbb{E}_\mathrm{P}(L \circ f - L \circ f^*_{L,\mathrm{P}}) + 4(c\, s)^q
\end{aligned}$$

for all $s > 0$, where in the last step we used Theorem 2.31 and the margin-exponent inequality. Optimizing over s by Lemma A.1.5 together with the estimate $(q + 1)2^{1/(q+1)}q^{-q/(q+1)} \le 3$ now yields the assertion. □

By combining the variance bound above with the analysis of Chapter 7, we can now formulate an oracle inequality for SVMs using Gaussian kernels.

Theorem 8.25 (Improved oracle inequality for Gaussian kernels). *Let* P *be a distribution on* $\mathbb{R}^d \times \{-1, 1\}$ *that has margin-noise exponent* $\beta \in (0, \infty)$ *and noise exponent* $q \in [0, \infty]$ *and whose* P_X *has tail exponent* $\tau \in (0, \infty]$. *Then, for all* $\varepsilon > 0$ *and all* $d/(d + \tau) < p < 1$, *there exists a constant* $K \ge 1$ *such that, for all fixed* $\varrho \ge 1$, $n \ge 1$, $\lambda \in (0, 1]$, *and* $\gamma \in (0, 1]$, *the SVM using the hinge loss* L *and a Gaussian RBF kernel with parameter* γ *satisfies*

$$\mathcal{R}_{L,\mathrm{P}}(\widehat{f}_{\mathrm{D},\lambda,\gamma}) - \mathcal{R}^*_{L,\mathrm{P}} \le K\left(\lambda^{\frac{\tau}{d+\tau}}\gamma^{-\frac{d\tau}{d+\tau}} + \gamma^\beta + \varrho\left(n\lambda^p\gamma^{(1-p)(1+\varepsilon)d}\right)^{-\frac{q+1}{q+2-p}}\right)$$

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

Proof. By Theorem 8.24, we have a variance bound of the form (7.36) for $\vartheta = q/(q+1)$, and the supremum bound (7.35) is obviously satisfied for $B = 2$. In addition, Theorem 7.34 together with Corollary 7.31 yields a bound (7.48) on the entropy numbers for the constant

$$a := c_{\varepsilon,p}\, \gamma^{-\frac{(1-p)(1+\varepsilon)d}{2p}},$$

where $d/(d+\tau) < p < 1$, $\varepsilon > 0$, and $c_{\varepsilon,p}$ is a constant only depending on ε and p. Finally, we set $f_0 := f^*$, where f^* is the function Theorem 8.18 provides. For $B_0 := 2$, Theorem 7.23 then yields the assertion. □

With the help of the oracle inequality above, we can now establish learning rates for SVMs using Gaussian kernels. For brevity's sake, we only mention the learning rates for the TV-SVM, but it should be clear from our previous considerations on this method that these learning rates match the fastest learning rates one could derive from Theorem 8.25.

Theorem 8.26 (Learning rates using Gaussian kernels). *Let $\Lambda := (\Lambda_n)$ and $\Gamma := (\Gamma_n)$ be sequences of finite subsets $\Lambda_n, \Gamma_n \subset (0,1]$ such that Λ_n is an n^{-1}-net of $(0,1]$ and Γ_n is an $n^{-1/d}$-net of $(0,1]$, respectively. Assume that the cardinalities of Λ_n and Γ_n grow polynomially in n. Then the TV-SVM with respect to Λ and Γ is universally consistent. Moreover, for distributions P on $\mathbb{R}^d \times \{-1,1\}$ that have margin-noise exponent $\beta \in (0,\infty)$ and noise exponent $q \in [0,\infty]$, and whose P_X have tail exponent $\tau \in (0,\infty]$, the TV-SVM learns with rate*

$$n^{-\frac{\beta\tau(d+\tau)(q+1)}{\beta\tau^2(q+2)+d(d\tau+\beta d+2\beta\tau+\tau^2)(q+1)}+\rho},$$

where $\rho > 0$ is an arbitrarily small number.

Proof. The consistency was already established in Theorem 8.21. Moreover, analogously to the proof of Theorem 6.31 and Theorem 7.24, we see that a simple union bound together with Theorem 8.25 shows for $m := \lfloor n/2 \rfloor + 1$ that

$$\mathcal{R}_{L,\mathrm{P}}(\widehat{f}_{\mathrm{D}_1,\lambda,\gamma}) - \mathcal{R}^*_{L,\mathrm{P}} \le K\left(\lambda^{\frac{\tau}{d+\tau}}\gamma^{-\frac{d\tau}{d+\tau}} + \gamma^{\beta} + \varrho\left(m\lambda^p\gamma^{(1-p)(1+\varepsilon)d}\right)^{-\frac{q+1}{q+2-p}}\right)$$

holds with probability P^m not less than $1 - 3|\Lambda_n| \cdot |\Gamma_n| e^{-\varrho}$ for all $\lambda \in \Lambda_n$ and $\gamma \in \Gamma_n$ simultaneously. As in the proof of Theorem 7.24, we then use Theorem 7.2 to deal with the parameter selection step, and combining both shows that with probability P^n not less than $1 - e^{-2\varrho}$ we have

$$\begin{aligned}
&\mathcal{R}_{L_{\text{class}},\mathrm{P}}(f_{\mathrm{D}_1,\lambda_{D_2},\gamma_{D_2}}) - \mathcal{R}^*_{L_{\text{class}},\mathrm{P}} \\
&\le 12K \inf_{\lambda\in\Lambda_n,\gamma\in\Gamma_n}\left(\lambda^{\frac{\tau}{d+\tau}}\gamma^{-\frac{d\tau}{d+\tau}} + \gamma^{\beta} + \varrho_n\left(n\lambda^p\gamma^{(1-p)(1+\varepsilon)d}\right)^{-\frac{q+1}{q+2-p}}\right) \\
&\quad + \tilde{K}\left(\frac{\left(\varrho + \ln(1+|\Lambda_n|\cdot|\Gamma_n|)\right)}{n}\right)^{\frac{q+1}{q+2}}, \qquad (8.17)
\end{aligned}$$

where $\varrho_n := \bigl(\varrho + \ln(1+3|\Lambda_n| \cdot |\Gamma_n|)\bigr)$ and $\tilde{K}$ is a suitable constant only depending on q and the constant c appearing in (8.16). Moreover, for all $\lambda \in \Lambda_n$ and $\gamma \in \Gamma_n$, we have

$$\left(\frac{\bigl(\varrho+\ln(1+|\Lambda_n|\cdot|\Gamma_n|)\bigr)}{n}\right)^{\frac{q+1}{q+2}} \leq \varrho_n \bigl(n\lambda^p \gamma^{(1-p)(1+\varepsilon)d}\bigr)^{-\frac{q+1}{q+2-p}},$$

and hence we may omit the last term in (8.17) if we replace $12K$ by $12K + \tilde{K}$. Repeating the proof of Corollary 8.19 and choosing p and ε sufficiently close to $d/(d+\tau)$ and 0, respectively, then yields the assertion after some simple yet tedious calculations. □

In order to illustrate the learning rates above, we assume in the following that P_X has tail exponent $\tau = \infty$. In this case, the learning rate reduces to

$$n^{-\frac{\beta(q+1)}{\beta(q+2)+d(q+1)}+\rho}, \tag{8.18}$$

where $\rho > 0$ is an arbitrarily small number. Motivated by the examples in Section 8.2, we first assume that P has margin exponent $\alpha := 1$. Moreover, we assume a moderate noise exponent $q := 1$, and by setting $\gamma := 1$ we also assume a moderate control of the noise by the distance to the decision boundary. By Lemma 8.17, these assumptions yield a margin-noise exponent $\beta = 2$, and hence (8.18) reduces to

$$n^{-\frac{2}{3+d}+\rho}.$$

Obviously, this rate is never faster than $n^{-1/2}$, and for high input dimensions it is actually substantially worse.

Let us now consider a different scenario that is less dimension dependent. To this end, we assume that there exists version η of the posterior probability such that the associated distance to the decision boundary Δ controls the noise by the exponent $\gamma \in [0,\infty)$. Lemma 8.23 then shows that $\beta = \gamma(q+1)$ and hence (8.18) reduces to

$$n^{-\frac{\gamma(q+1)}{\gamma(q+2)+d}+\rho}.$$

For large q or γ, this rate is obviously rather insensitive to the input dimension, and in particular for $q \to \infty$ we obtain rates that are close to the rate n^{-1}. Moreover, note that large q reflects both a small amount of noise and, by Lemma 8.23, a low concentration of P_X near the decision boundary.

8.4 Sparseness of SVMs Using the Hinge Loss

We have seen in Theorem 5.5 that using a convex loss function there exists a unique SVM solution $f_{\mathrm{D},\lambda}$, which, in addition, is of the form

$$f_{\mathrm{D},\lambda} = \sum_{i=1}^{n} \alpha_i k(\,\cdot\,, x_i)\,, \tag{8.19}$$

where $\alpha = (\alpha_1, \ldots, \alpha_n) \in \mathbb{R}^n$ is a suitable vector of coefficients, and $D = ((x_1, y_1), \ldots, (x_n, y_n))$. Obviously, only **support vectors**, i.e., samples x_i whose coefficients α_i are non-zero, need to be considered when evaluating $f_{\mathrm{D},\lambda}(x)$ by (8.19), and hence the number of support vectors has a direct influence on the time required to evaluate $f_{\mathrm{D},\lambda}(x)$. Moreover, we will see in Section 11.2 that this number also has a substantial influence on the training time. Consequently, it is important to know whether we can expect **sparse** decision functions, i.e., decision functions for which not all samples are support vectors. The goal of this section is to present some results that estimate the typical number of support vectors when using the hinge loss. In the following section, we will then extend these considerations to general convex, classification calibrated, margin-based loss functions.

Let us begin by considering the (quadratic) optimization problem

$$\begin{array}{lll} \text{minimize} & \lambda\|f\|_H^2 + \frac{1}{n}\sum\limits_{i=1}^{n} \xi_i & \text{for } f \in H,\, \xi \in \mathbb{R}^n \\ \text{subject to} & y_i f(x_i) \geq 1 - \xi_i, & i = 1, \ldots, n \\ & \xi_i \geq 0, & i = 1, \ldots, n\,. \end{array} \tag{8.20}$$

It is obvious that a pair $(f^*, \xi^*) \in H \times \mathbb{R}^n$ with $\xi_i^* = \max\{0, 1 - y_i f^*(x_i)\}$ is a solution of (8.20) if and only if $f^* = f_{\mathrm{D},\lambda}$. Consequently, one can solve (8.20) in order to find the SVM decision function. We will see in Chapter 11 that in practice this quadratic optimization problem is usually solved by considering the dual problem

$$\begin{array}{lll} \text{maximize} & \sum\limits_{i=1}^{n} \alpha_i - \frac{1}{4\lambda}\sum\limits_{i,j=1}^{n} y_i y_j \alpha_i \alpha_j k(x_i, x_j) & \text{for } \alpha \in \mathbb{R}^n \\ \text{subject to} & 0 \leq \alpha_i \leq \frac{1}{n}, & i = 1, \ldots, n\,, \end{array} \tag{8.21}$$

where we note that in Example 11.3 the primal problem will first be rescaled. Note that this rescaling results in the differently scaled but otherwise identical dual problem (11.15), and hence we can consider (8.21) instead of the problem (11.15). In order to establish a lower bound on the number of support vectors, we now have to briefly discuss the relation between (8.20) and (8.21). To this end, we write

$$f^{(\alpha)} := \frac{1}{2\lambda} \sum_{i=1}^{n} y_i \alpha_i k(\,\cdot\,, x_i) \tag{8.22}$$

and $\xi_i^{(\alpha)} := \max\{0, 1 - y_i f^{(\alpha)}(x_i)\}$, where $\alpha \in \mathbb{R}^n$ is an arbitrary vector. Furthermore, we define

$$\begin{aligned}\operatorname{gap}(\alpha) &:= \lambda\|f^{(\alpha)}\|_H^2 + \frac{1}{n}\sum_{i=1}^n \xi_i^{(\alpha)} - \sum_{i=1}^n \alpha_i + \frac{1}{4\lambda}\sum_{i,j=1}^n y_i y_j \alpha_i \alpha_j k(x_i, x_j) \\ &= 2\lambda\|f^{(\alpha)}\|_H^2 + \frac{1}{n}\sum_{i=1}^n \xi_i^{(\alpha)} - \sum_{i=1}^n \alpha_i \end{aligned} \tag{8.23}$$

for the difference between the value of the objective function of (8.20) and (8.21). Theorem A.6.28 together with Example 11.3 shows that if $\alpha^* \in \mathbb{R}^n$ is a solution of (8.21), then $(f^{(\alpha^*)}, \xi^{(\alpha^*)})$ is a solution of (8.20), i.e.,

$$f_{\mathrm{D},\lambda} = \frac{1}{2\lambda}\sum_{i=1}^n y_i \alpha_i^* k(\,\cdot\,, x_i)\,, \tag{8.24}$$

and we have $\operatorname{gap}(\alpha^*) = 0$. In other words, the SVM solution $f_{\mathrm{D},\lambda}$ can be obtained by solving the dual problem (8.21) and substituting the corresponding solution α^* into (8.22). Unfortunately, however, it is almost always impossible to find an *exact* solution α^* of (8.21). Consequently, let us assume that we have an $\alpha \in \mathbb{R}^n$ that is feasible for (8.21), i.e., $\alpha_i \in [0, 1/n]$ for all $i = 1, \ldots, n$. From the already mentioned fact that an exact solution α^* satisfies $\operatorname{gap}(\alpha^*) = 0$, we then conclude that

$$\begin{aligned}\operatorname{gap}(\alpha) &\geq \lambda\|f^{(\alpha)}\|_H^2 + \frac{1}{n}\sum_{i=1}^n \xi_i^{(\alpha)} - \sum_{i=1}^n \alpha_i^* + \frac{1}{4\lambda}\sum_{i,j=1}^n y_i y_j \alpha_i^* \alpha_j^* k(x_i, x_j) \\ &= \lambda\|f^{(\alpha)}\|_H^2 + \frac{1}{n}\sum_{i=1}^n \xi_i^{(\alpha)} - \lambda\|f^{(\alpha^*)}\|_H^2 - \frac{1}{n}\sum_{i=1}^n \xi_i^{(\alpha^*)}\,. \end{aligned}$$

Using the definitions of $f^{(\alpha)}$, $\xi^{(\alpha)}$, and $f^{(\alpha^*)} = f_{\mathrm{D},\lambda}$, we thus obtain

$$\lambda\|f^{(\alpha)}\|_H^2 + \mathcal{R}_{L,\mathrm{D}}(f^{(\alpha)}) \;\leq\; \inf_{f \in H} \lambda\|f\|_H^2 + \mathcal{R}_{L,\mathrm{D}}(f) + \operatorname{gap}(\alpha)$$

for all $\alpha \in [0, 1/n]^n$. Consequently, if we have an $\alpha \in [0, 1/n]^n$ with $\operatorname{gap}(\alpha) \leq \epsilon$, where $\epsilon > 0$ is some predefined accuracy, then $f^{(\alpha)}$ is a decision function satisfying the ϵ-CR-ERM inequality (7.31), and hence the statistical analysis of Section 7.4 can be applied.

Our next goal is to estimate the number of non-zero coefficients of $f^{(\alpha)}$ when represented by (8.22). We begin with the following proposition.

Proposition 8.27 (Deterministic lower bound on the sparseness). *Let X be a non-empty set, k be a kernel on X, $D \in (X \times Y)^n$ be an arbitrary sample set, and $\alpha = (\alpha_1, \ldots, \alpha_n) \in \mathbb{R}^n$. If α is feasible, i.e., $0 \leq \alpha_i \leq 1/n$ for all $i = 1, \ldots, n$, then we have*

$$\frac{|\{i : \alpha_i \neq 0\}|}{n} \;\geq\; 2\lambda\|f^{(\alpha)}\|_H^2 + \mathcal{R}_{L,\mathrm{D}}(f^{(\alpha)}) - \operatorname{gap}(\alpha)\,.$$

Proof. By (8.23) and the definition of $\xi^{(\alpha)}$, we obtain

$$\frac{|\{i : \alpha_i \neq 0\}|}{n} \;=\; \sum_{\alpha_i \neq 0} \frac{1}{n} \;\geq\; \sum_{i=1}^{n} \alpha_i \;=\; 2\lambda \|f^{(\alpha)}\|_H^2 + \mathcal{R}_{L,\mathrm{D}}(f^{(\alpha)}) - \operatorname{gap}(\alpha)\,.$$

□

Our next goal is to establish a probabilistic bound on the number of non-zero coefficients. To this end, we simplify the presentation by considering only *exact* solutions α^* and $f_{\mathrm{D},\lambda}$, though similar results also hold for approximate solutions. Note that in this case we have $\operatorname{gap}(\alpha^*) = 0$, and hence Proposition 8.27 yields

$$\frac{|\{i : \alpha_i^* \neq 0\}|}{n} \;\geq\; 2\lambda \|f_{\mathrm{D},\lambda}\|_H^2 + \mathcal{R}_{L,\mathrm{D}}(f_{\mathrm{D},\lambda})\,.$$

Consequently, we need a lower bound on the right-hand side of this estimate. A relatively simple lower bound is presented in the following proposition.

Proposition 8.28 (Probabilistic lower bound on the sparseness). *Let X be a compact metric space, H be the RKHS of a continuous kernel k on X with $\|k\|_\infty \leq 1$, and P be a probability measure on $X \times Y$. Then, for fixed $\lambda \in (0,1]$, $n \geq 1$, $\varepsilon > 0$, and $\tau > 0$, we have with probability P^n not less than $1 - e^{-\tau}$ that*

$$\lambda \|f_{\mathrm{D},\lambda}\|_H^2 + \mathcal{R}_{L,\mathrm{D}}(f_{\mathrm{D},\lambda}) \;>\; \lambda \|f_{\mathrm{P},\lambda}\|_H^2 + \mathcal{R}_{L,\mathrm{P}}(f_{\mathrm{P},\lambda}) - 2\varepsilon - \sqrt{\frac{2\tau + 2\ln \mathcal{N}(B_H, \|\cdot\|_\infty, \lambda^{1/2}\varepsilon)}{\lambda n}}\,.$$

Proof. We have $\lambda \|f_{\mathrm{P},\lambda}\|_H^2 + \mathcal{R}_{L,\mathrm{P}}(f_{\mathrm{P},\lambda}) \leq \lambda \|f_{\mathrm{D},\lambda}\|_H^2 + \mathcal{R}_{L,\mathrm{P}}(f_{\mathrm{D},\lambda})$ by the definition of $f_{\mathrm{P},\lambda}$, and hence we obtain

$$\begin{aligned}
&\lambda \|f_{\mathrm{P},\lambda}\|_H^2 + \mathcal{R}_{L,\mathrm{P}}(f_{\mathrm{P},\lambda}) - \lambda \|f_{\mathrm{D},\lambda}\|_H^2 - \mathcal{R}_{L,\mathrm{D}}(f_{\mathrm{D},\lambda}) \\
&\leq \mathcal{R}_{L,\mathrm{P}}(f_{\mathrm{D},\lambda}) - \mathcal{R}_{L,\mathrm{D}}(f_{\mathrm{D},\lambda}) \\
&\leq \sup_{f \in \lambda^{-1/2} B_H} \bigl(\mathcal{R}_{L,\mathrm{P}}(f) - \mathcal{R}_{L,\mathrm{D}}(f)\bigr) \\
&\leq \sup_{g \in \mathcal{F}_\varepsilon} \bigl(\mathcal{R}_{L,\mathrm{P}}(g) - \mathcal{R}_{L,\mathrm{D}}(g)\bigr) + 2\varepsilon\,,
\end{aligned}$$

where $\mathcal{F}_\varepsilon \subset \lambda^{-1/2} B_H$ is assumed to be an ε-net of $\lambda^{-1/2} B_H$ having cardinality $\mathcal{N}(\lambda^{-1/2} B_H, \|\cdot\|_\infty, \varepsilon) = \mathcal{N}(B_H, \|\cdot\|_\infty, \lambda^{1/2}\varepsilon)$. Moreover, for $g \in \lambda^{-1/2} B_H$, we have $\|g\|_\infty \leq \lambda^{-1/2}$, and hence we find

$$L(y, g(x)) = \max\{0, 1 - y g(x)\} \leq 1 + g(x) \leq 2\lambda^{-1/2}\,, \qquad y = \pm 1,\, x \in X.$$

By a union bound and Hoeffding's inequality, we thus obtain

$$\mathrm{P}^n\left(D \in (X \times Y)^n : \sup_{g \in \mathcal{F}_\varepsilon} \bigl(\mathcal{R}_{L,\mathrm{P}}(g) - \mathcal{R}_{L,\mathrm{D}}(g)\bigr) < \sqrt{\frac{2\tau}{\lambda n}}\right) \geq 1 - |\mathcal{F}_\varepsilon|\, e^{-\tau}\,.$$

Combining this estimate with the first one then yields the assertion. □

In order to illustrate how the results above can be combined, let us assume for a moment that we use a *fixed* RKHS H whose covering numbers satisfy the usual assumption

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

where $a, p > 0$ are some constants independent of ε. For $\lambda \in (0,1]$, $\tau > 0$, $n \geq 1$, and $\varepsilon := (\frac{p}{2})^{1/(1+p)}(\frac{2a}{n})^{1/(2+2p)}\lambda^{-1/2}$, Proposition 8.28 then shows that

$$\lambda\|f_{\mathrm{D},\lambda}\|_H^2 + \mathcal{R}_{L,\mathrm{D}}(f_{\mathrm{D},\lambda}) > \mathcal{R}^*_{L,\mathrm{P},H} - 4\Big(\frac{a}{\lambda^{1+p}n}\Big)^{\frac{1}{2+2p}} - \sqrt{\frac{2\tau}{\lambda n}}$$

holds with probability P^n not less than $1-e^{-\tau}$. Combining this estimate with Proposition 8.27, we find that

$$\frac{|\{i : \alpha_i^*(D) \neq 0\}|}{n} \;\geq\; \mathcal{R}^*_{L,\mathrm{P},H} - 4\Big(\frac{a}{\lambda^{1+p}n}\Big)^{\frac{1}{2+2p}} - \sqrt{\frac{2\tau}{\lambda n}}$$

holds with probability P^n not less than $1 - e^{-\tau}$, where $(\alpha_1^*(D), \ldots, \alpha_n^*(D))$ is an exact solution of the dual problem (8.21) defined by the sample set $D = ((x_1, y_1), \ldots, (x_n, y_n))$. In particular, if we use an *a priori* chosen regularization sequence $(\lambda_n) \subset (0,1]$ satisfying both $\lambda_n \to 0$ and $\lambda_n^{1+p}n \to \infty$ and an RKHS H such that $\mathcal{R}^*_{L,\mathrm{P},H} = \mathcal{R}^*_{L,\mathrm{P}}$, then the SVM is (classification) consistent as we have seen around (6.22), *and* for all $\varepsilon > 0$ we have

$$\lim_{n\to\infty} \mathrm{P}^n\left(D \in (X \times Y)^n : \frac{|\{i : \alpha_i^*(D) \neq 0\}|}{n} \;\geq\; \mathcal{R}^*_{L,\mathrm{P}} - \varepsilon\right) = 1\,. \tag{8.25}$$

In other words, the number of support vectors tends to be not smaller than $\mathcal{R}^*_{L,\mathrm{P}}$. Moreover, recalling the calculations in Example 2.4 and the proof of Theorem 2.31, we find $\mathcal{R}^*_{L,\mathrm{P}} = 2\mathcal{R}^*_{L_{\mathrm{class}},\mathrm{P}}$, i.e., asymptotically the number of support vectors is not smaller than twice the Bayes classification risk. Finally, one can generalize this result to *data-dependent* methods of choosing λ such as the one considered in Section 6.5. The details can be found in Exercise 8.7.

Note that the results above describe the number of support vectors when we use the dual problem (8.21) for finding $f_{\mathrm{D},\lambda}$. However, we will see in the next section that if the RKHS H is universal and P_X is an atom-free distribution, then we almost surely have a unique representation (8.19), and consequently the above lower bounds on the number of support vectors hold for *every* algorithm producing $f_{\mathrm{D},\lambda}$. On the other hand, for *finite-dimensional* RKHSs H, the number of support vectors does depend on the algorithm. We refer to Exercise 8.6 for details.

8.5 Classifying with other Margin-Based Losses (*)

Although the hinge loss is the most commonly used loss for binary classification with SVMs, it is by no means the only choice. Indeed we have seen in

Section 3.4 that there are various other (convex) margin-based loss functions that are calibrated to the classification loss. Moreover, by Theorem 3.34, these losses are automatically uniformly classification calibrated and hence they are excellent alternatives if the sole objective is classification. However, some of these losses possess nice additional features that are important for some applications. For example, the (truncated) least squares loss and the logistic loss enable us to estimate posterior probabilities (see Example 3.66 for the latter), and hence these losses can be interesting whenever such an estimate is relevant for the application at hand. In addition, different loss functions lead to different optimization problems, and it is possible that some of these optimization problems promise algorithmic advantages over the one associated with the hinge loss. Consequently, there are situations in which the hinge loss may *not* be the first choice and alternatives are desired. In this section, we investigate some features of such alternatives so that an informed decision can be made. In particular, we will revisit calibration inequalities, establish a lower bound on the number of support vectors, and describe a relation between sparseness and the possibility of estimating the posterior probability.

Let us begin by recalling Theorem 3.36, which showed that a convex margin-based loss L is classification calibrated if and only if its representing function $\varphi : \mathbb{R} \to [0,\infty)$ is differentiable at 0 with $\varphi'(0) < 0$. Moreover, Theorem 3.34 showed that in this case L is actually uniformly classification calibrated and that the uniform calibration function is given by (3.41). Now assume for simplicity that this calibration function satisfies $\delta^{**}_{\max}(\varepsilon, \mathcal{Q}_Y) \geq c_p \varepsilon^p$ for some constants $c_p > 0$, $p \geq 1$, and all $\varepsilon \in [0,1]$. Then Theorem 3.22 yields

$$\mathcal{R}_{L_{\text{class}},\mathrm{P}}(f) - \mathcal{R}^*_{L_{\text{class}},\mathrm{P}} \;\leq\; c_p^{-1/p} \bigl(\mathcal{R}_{L,\mathrm{P}}(f) - \mathcal{R}^*_{L,\mathrm{P}}\bigr)^{1/p} \tag{8.26}$$

for all distributions P on $X \times Y$ and all $f : X \to \mathbb{R}$. The next result shows that (8.26) can be improved if $p > 1$ and P has a non-trivial noise exponent.

Theorem 8.29 (Inequality for classification calibrated losses). *Let L be a convex, margin-based, and classification calibrated loss whose uniform calibration function satisfies $\delta^{**}_{\max}(\varepsilon, \mathcal{Q}_Y) \geq c_p \varepsilon^p$ for some constants $c_p > 0$, $p > 1$, and all $\varepsilon \in [0,1]$. Moreover, let* P *be a distribution on $X \times Y$ that has some noise exponent $q \in [0,\infty]$. Then, for all measurable $f : X \to \mathbb{R}$, we have*

$$\mathcal{R}_{L_{\text{class}},\mathrm{P}}(f) - \mathcal{R}^*_{L_{\text{class}},\mathrm{P}} \;\leq\; 2 c_p^{-\frac{q+1}{q+p}} c^{\frac{q(p-1)}{q+p}} \bigl(\mathcal{R}_{L,\mathrm{P}}(f) - \mathcal{R}^*_{L,\mathrm{P}}\bigr)^{\frac{q+1}{q+p}},$$

where c is the constant appearing in the noise exponent inequality (8.16).

Proof. We have seen in Remark 3.35 that $L_{\mathrm{P}} : X \times \mathbb{R} \to [0,\infty)$ defined by

$$L_{\mathrm{P}}(x,t) := \bigl|2\eta(x) - 1\bigr| \cdot \mathbf{1}_{(-\infty,0]}\bigl((2\eta(x) - 1)\operatorname{sign} t\bigr), \qquad x \in X, t \in \mathbb{R},$$

is a detection loss with respect to $A := \{(x,t) \in X \times \mathbb{R} : (2\eta(x) - 1)\operatorname{sign} t \leq 0\}$ and $h(x) = |2\eta(x) - 1|$, $x \in X$, where $\eta(x) := \mathrm{P}(y = 1|x)$. Moreover, (3.9) shows for the inner excess risks that

$$\mathcal{C}_{L_{\mathrm{P}},x}(t) - \mathcal{C}^*_{L_{\mathrm{P}},x} \;=\; \mathcal{C}_{L_{\mathrm{class}},\eta(x)}(t) - \mathcal{C}^*_{L_{\mathrm{class}},\eta(x)}\,, \qquad x \in X,\, t \in \mathbb{R},$$

i.e., L_{P} describes the classification goal for the distribution P. In particular, we have $\delta_{\max,L_{\mathrm{P}},L}(\varepsilon, \mathrm{P}(\,\cdot\,|x), x) = \delta_{\max,L_{\mathrm{class}},L}(\varepsilon, \mathrm{P}(\,\cdot\,|x))$ for all $x \in X$, $\varepsilon \in [0,\infty]$, and hence we obtain

$$\delta_{\max,L_{\mathrm{P}},L}\bigl(\varepsilon, \mathrm{P}(\,\cdot\,|x), x\bigr) \;\geq\; \delta^{**}_{\max,L_{\mathrm{class}},L}(\varepsilon, \mathcal{Q}_Y) \;\geq\; c_p \varepsilon^p$$

for all $x \in X$ and $\varepsilon \in [0,1]$. Since $\|h\|_\infty \leq 1$, we conclude that

$$\begin{aligned} B(s) &:= \bigl\{x \in X : \delta_{\max,L_{\mathrm{P}},L}\bigl(h(x), \mathrm{P}(\,\cdot\,|x), x\bigr) < s\,h(x)\bigr\} \\ &\subset \bigl\{x \in X : |2\eta(x) - 1| < (s/c_p)^{1/(p-1)}\bigr\} \end{aligned}$$

for all $s > 0$. Using the noise exponent, we thus obtain

$$\begin{aligned} \int_X \mathbf{1}_{B(s)} h \, d\mathrm{P}_X &\leq (s/c_p)^{1/(p-1)} \mathrm{P}_X\bigl(\bigl\{x \in X : |2\eta(x) - 1| < (s/c_p)^{1/(p-1)}\bigr\}\bigr) \\ &\leq \Bigl(c^{\frac{q(p-1)}{q+1}} \cdot \frac{s}{c_p}\Bigr)^{\frac{q+1}{p-1}} \end{aligned}$$

for all $s > 0$. Now the assertion follows from Theorem 3.28. □

In Table 3.1, we see that the (truncated) least squares L loss satisfies $\delta^{**}_{\max}(\varepsilon, \mathcal{Q}_Y) \geq \varepsilon^2$, and hence the inequality above reduces to

$$\mathcal{R}_{L_{\mathrm{class}},\mathrm{P}}(f) - \mathcal{R}^*_{L_{\mathrm{class}},\mathrm{P}} \;\leq\; 2c^{\frac{q}{q+2}} \bigl(\mathcal{R}_{L,\mathrm{P}}(f) - \mathcal{R}^*_{L,\mathrm{P}}\bigr)^{\frac{q+1}{q+2}}\,. \tag{8.27}$$

Moreover, for the logistic loss for classification, we have $\delta^{**}_{\max}(\varepsilon, \mathcal{Q}_Y) \geq \varepsilon^2/2$ and hence it satisfies (8.27) if we replace the factor 2 by 4. Finally, note that both factors can be improved by a more careful analysis in the proof of Theorem 3.28.

Let us now consider the number of support vectors we may expect using different convex margin-based loss functions. To this end, recall again that Theorem 5.5 showed that the unique SVM solution $f_{\mathrm{D},\lambda}$ is of the form

$$f_{\mathrm{D},\lambda} = \sum_{i=1}^n \alpha_i k(\,\cdot\,, x_i)\,, \tag{8.28}$$

where $\alpha = (\alpha_1, \ldots, \alpha_n) \in \mathbb{R}^n$ is a suitable vector of coefficients and $D = ((x_1, y_1), \ldots, (x_n, y_n))$. Let us assume for a moment that we have $x_i \neq x_j$ for all $i \neq j$, where we note that this is almost surely satisfied if D is sampled from a distribution P with $\mathrm{P}(\{x\}) = 0$ for all $x \in X$. In addition, we assume throughout this section that the kernel k is strictly positive definite, where we recall that this assumption is necessary (but in general not sufficient) for universal consistency by Corollary 5.34. From linear algebra, we then know that the kernel matrix $K := (k(x_j, x_i))_{i,j=1}^n$ has P^n-almost surely full rank, and by considering the system of linear equations

$$f_{\mathrm{D},\lambda}(x_j) = \sum_{i=1}^{n} \alpha_i k(x_j, x_i)\,, \qquad j = 1, \ldots, n\,,$$

we see that the representation (8.28) is P^n-almost surely unique, i.e., there exists exactly one $\alpha \in \mathbb{R}^n$ satisfying (8.28). In the following, we thus write

$$\#SV(f_{\mathrm{D},\lambda}) := |\{i : \alpha_i \neq 0\}|\,,$$

where we note that this quantity is only P^n-almost surely defined.

Let us now assume that L is a convex margin-based loss with representing function φ, i.e., $L(y,t) = \varphi(yt)$, $y = \pm 1$, $t \in \mathbb{R}$. Since the subdifferential of L is given by $\partial L(y,\cdot) = y\partial\varphi(y\,\cdot\,)$ for $y = \pm 1$, Corollary 5.12 then shows that the almost surely uniquely defined coefficients α_i in (8.28) satisfy

$$\alpha_i \in -\frac{y_i}{2\lambda n}\partial\varphi\big(y_i f_{\mathrm{D},\lambda}(x_i)\big)\,, \qquad i = 1, \ldots, n\,. \tag{8.29}$$

In particular, if φ is differentiable, we have

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

Now (8.29) shows that x_i must be a support vector if $0 \notin \partial\varphi(y_i f_{\mathrm{D},\lambda}(x_i))$. This immediately leads to the following preliminary result.

Proposition 8.30 (No sparseness without global minimum). *Let L be a convex, margin-based loss with representing function φ. Moreover, let k be a strictly positive definite kernel on X and P be a distribution on $X \times Y$ such that $\mathrm{P}(\{x\}) = 0$ for all $x \in X$. If φ does not have a global minimum, then for P^n-almost all $D \in (X \times Y)^n$ we have $\#SV(f_{\mathrm{D},\lambda}) = n$ for all $\lambda > 0$.*

Proof. If φ does not have a global minimum, then $0 \notin \partial\varphi(t)$ for all $t \in \mathbb{R}$ and hence (8.29) yields the assertion. □

Note that the logistic loss for classification satisfies the assumptions of the preceding proposition and hence we can in general not expect to obtain sparse decision functions when using this loss.

Because of Proposition 8.30, it remains to investigate convex margin-based losses L whose representing functions φ do have at least one global minimum. Note that by Lemma 2.23 the latter assumption is satisfied if and only if L can be clipped, and hence these loss functions enjoy the advanced oracle inequalities of Section 7.4. Now recall that Lemma 3.64 together with Lemma 3.60 showed that such losses are also self-calibrated for *all* distributions Q on $\{-1, 1\}$. Consequently, if D is a training set such that $\mathcal{R}_{L,\mathrm{P}}(f_{\mathrm{D},\lambda})$ is close to $\mathcal{R}^*_{L,\mathrm{P}}$, then $f_{\mathrm{D},\lambda}$ is close to the set-valued function $x \mapsto \mathcal{M}_{L,\eta(x)}(0^+)$ of exact minimizers, where, as usual, $\eta(x) := \mathrm{P}(y = 1|x)$. This suggests that x_i must be a support vector of the representation above whenever $0 \notin \partial\varphi(y_i\mathcal{M}_{L,\eta(x_i)}(0^+))$. Our next goal is to verify this intuition. We begin with a simple lemma that collects some useful properties of $\mathcal{M}_{L,\eta}(0^+)$.

Lemma 8.31. *Let L be a convex, classification calibrated, and margin-based loss that can be clipped. Then, for all $t \in \mathbb{R}$ and $\eta \in [0,1]$, the following statements are true:*

i) The set $\mathcal{M}_{L,\eta}(0^+)$ is a bounded, closed interval whenever $\eta \in (0,1)$.
ii) If $\partial L(1,t) \cap [0,\infty) \neq \emptyset$, then we have $t > 0$.
iii) We have $0 \notin \partial L(1,t) \cap \partial L(-1,t)$.
iv) If $t \in \mathcal{M}_{L,\eta}(0^+)$, then there exist $s^+ \in \partial L(1,t) \cap (-\infty, 0]$ and $s^- \in \partial L(-1,t) \cap [0,\infty)$ such that $\eta s^+ + (1-\eta)s^- = 0$.
v) If $t \in \mathcal{M}_{L,\eta}(0^+)$, we have $\min \partial \mathcal{C}_{L,\eta'}(t) < 0$ for all $\eta' > \eta$.
vi) The set-valued map $\eta \mapsto \mathcal{M}_{L,\eta}(0^+)$ is a monotone operator, i.e., we have $\sup \mathcal{M}_{L,\eta}(0^+) \leq \inf \mathcal{M}_{L,\eta'}(0^+)$ for all $\eta' \in [0,1]$ with $\eta' > \eta$.

Proof. i). Let $\varphi : \mathbb{R} \to [0,\infty)$ be the function representing L. By Theorem 3.36, we know $L'(1,0) = \varphi'(0) < 0$ and, since φ is convex, we conclude that $\lim_{t\to-\infty} \varphi(t) = \infty$. For $\eta \in (0,1)$, the latter implies $\lim_{t\to\pm\infty} \mathcal{C}_{L,\eta}(t) = \infty$, and hence $\mathcal{M}_{L,\eta}(0^+)$ is bounded. The remaining assertions follow from the continuity and convexity of $t \mapsto \mathcal{C}_{L,\eta}(t)$.

ii). Assume we had $t \leq 0$. Then the monotonicity of subdifferentials together with Theorem 3.36 would yield $s \leq \varphi'(0) < 0$ for all $s \in \partial L(1,t)$.
iii). Assume that there exists a $t \in \mathbb{R}$ such that $0 \in \partial L(1,t) \cap \partial L(-1,t)$. By *ii)*, we then see that $0 \in \partial L(1,t)$ implies $t > 0$, while $0 \in \partial L(-1,t) = -\partial L(1,-t)$ implies $t < 0$.

iv). For a given $t \in \mathcal{M}_{L,\eta}(0^+)$, there exist an $s^+ \in \partial L(1,t)$ and an $s^- \in \partial L(-1,t)$ with $0 = \eta s^+ + (1-\eta)s^-$. If $\eta = 1$, we find $s^+ = 0$. By *ii)* this implies $t > 0$, and, since $s^- \in \partial L(-1,t) = -\partial L(1,-t)$, we thus have $s^- > 0$ by *ii)*. The case $\eta = 0$ can be treated analogously. Finally, if $0 < \eta < 1$, we have $(1-\eta)s^- = -\eta s^+$, which leads to either $s^- \leq 0 \leq s^+$ or $s^+ \leq 0 \leq s^-$. Moreover, *ii)* shows that $s^+ \geq 0$ implies $t > 0$ and that $s^- \leq 0$ implies $t < 0$, and hence we conclude that $s^+ \leq 0 \leq s^-$.

v). Without loss of generality, we may assume $\eta < 1$. Let us fix $s^+ \in \partial L(1,t)$ and $s^- \in \partial L(-1,t)$ according to *iv)*. Then we find $s^+ - s^- < 0$ by *iii)* and hence

$$s \;:=\; \eta' s^+ + (1-\eta')s^- \;<\; \eta s^+ + (1-\eta)s^- \;=\; 0\,.$$

Finally, the linearity of subdifferentials yields $s \in \partial \mathcal{C}_{L,\eta'}(t)$.

vi). Let $0 \leq \eta < \eta' \leq 1$ as well as $t \in \mathcal{M}_{L,\eta}(0^+)$ and $t' \in \mathcal{M}_{L,\eta'}(0^+)$. By *v)*, we find an $s \in \partial \mathcal{C}_{L,\eta'}(t)$ with $s < 0$. Then we obtain $t' \geq t$ since otherwise the monotonicity of subdifferentials implies $s' \leq s < 0$ for all $s' \in \partial \mathcal{C}_{L,\eta'}(t')$, which in turn contradicts $t' \in \mathcal{M}_{L,\eta'}(0^+)$. □

The next lemma shows that $0 \notin \partial\varphi(t)$ implies $0 \notin \partial\varphi(s)$ for all s that are sufficiently close to t. This fact will be crucial for verifying our intuition since in general we cannot guarantee $f_{\mathrm{D},\lambda} \in \mathcal{M}_{L,\eta(x)}(0^+)$.

Lemma 8.32. *Let L be a convex, classification calibrated, and margin-based loss that can be clipped and* P *be a distribution on $X \times Y$. For $\varepsilon \geq 0$, we define*

$$S_\varepsilon := \left\{(x,y) \in X \times Y : 0 \notin \partial L\big(y, \mathcal{M}_{L,\eta(x)}(0^+) + \varepsilon B_{\mathbb{R}}\big)\right\}.$$

Then we have $S_\varepsilon \subset S_{\varepsilon'}$ for all $\varepsilon > \varepsilon' \geq 0$. Moreover, we have

$$\bigcup_{\varepsilon>0} S_\varepsilon = \left\{(x,y) \in X \times Y : 0 \notin \partial L\big(y, \mathcal{M}_{L,\eta(x)}(0^+)\big)\right\}.$$

Proof. Since the first assertion is trivial, it suffices to prove $S_0 \subset \bigcup_{\varepsilon>0} S_\varepsilon$. Obviously, this follows once we have established

$$\bigcap_{\varepsilon>0} \bigcup_{\delta\in[-\varepsilon,\varepsilon]} \bigcup_{t\in\mathcal{M}_{L,\eta}(0^+)} \partial L(y, t+\delta) \subset \bigcup_{t\in\mathcal{M}_{L,\eta}(0^+)} \partial L(y,t) \tag{8.30}$$

for all $\eta \in [0,1]$, $y = \pm 1$. Let us fix an element s from the set on the left-hand side of (8.30). Then, for all $n \in \mathbb{N}$, there exist $\delta_n \in [-1/n, 1/n]$ and $t_n \in \mathcal{M}_{L,\eta}(0^+)$ with $s \in \partial L(y, t_n + \delta_n)$. If (t_n) is unbounded, we obtain $\eta \in \{0,1\}$ by *i)* of Lemma 8.31. Furthermore, in this case we find $t_n + \delta_n \in \mathcal{M}_{L,\eta}(0^+)$ for all sufficiently large n since $\mathcal{M}_{L,\eta}(0^+)$ is an interval by the convexity of L. Hence we have shown (8.30) in this case. On the other hand, if (t_n) is bounded, there exists a subsequence (t_{n_k}) of (t_n) converging to a $t_0 \in \mathbb{R}$, and since $\mathcal{M}_{L,\eta}(0^+)$ is closed, we find $t_0 \in \mathcal{M}_{L,\eta}(0^+)$. Now let us fix an $\varepsilon > 0$. By Proposition A.6.14, we find that

$$s \in \partial L(y, t_{n_k} + \delta_{n_k}) \subset \partial L(y, t_0) + \varepsilon B_{\mathbb{R}}$$

for a sufficiently large k. This yields

$$s \in \bigcap_{\varepsilon>0} \big(\partial L(y,t_0) + \varepsilon B_{\mathbb{R}}\big),$$

and thus we finally obtain $s \in \partial L(y,t_0)$ by the compactness of $\partial L(y,t_0)$. □

Before we establish a lower bound on the number of support vectors, we need another elementary lemma.

Lemma 8.33. *Let L be a convex, classification calibrated, and margin-based loss that can be clipped and* P *be a distribution on $X \times Y$. For a function $f : X \to \mathbb{R}$ and $\varepsilon > 0$, we write*

$$A(f,\varepsilon) := \left\{(x,y) \in X \times Y : \operatorname{dist}\big(f(x), \mathcal{M}_{L,\eta(x)}(0^+)\big) \geq \varepsilon\right\}.$$

Then, for all $\varepsilon > 0$, all $f, g \in \mathcal{L}_\infty(X)$ satisfying $\|f - g\|_\infty \leq \varepsilon$, and all $(x,y) \in S_{2\varepsilon} \setminus A(g,\varepsilon)$, we have

$$0 \notin \partial\varphi(yf(x)).$$

Proof. We fix an element $(x,y) \in S_{2\varepsilon}$ such that $(x,y) \notin A(g,\varepsilon)$. Then we have $\operatorname{dist}(g(x), \mathcal{M}_{L,\eta(x)}(0^+)) < \varepsilon$, and thus there exists a $t \in \mathcal{M}_{L,\eta(x)}(0^+)$ such that $|g(x) - t| < \varepsilon$. This implies $\operatorname{dist}(f(x), \mathcal{M}_{L,\eta(x)}(0^+)) \leq |f(x) - t| < 2\varepsilon$ by the triangle inequality, i.e., we have $f(x) \in \mathcal{M}_{L,\eta(x)}(0^+) + 2\varepsilon B_{\mathbb{R}}$. The definition of $S_{2\varepsilon}$ together with $\partial L(y,t) = y\partial\varphi(yt)$ then yields the assertion. □

With these preparations, we can now establish the announced lower bound on the number of support vectors.

Theorem 8.34 (Probabilistic lower bound on the sparseness). *Let L be a convex, classification calibrated, and margin-based loss whose representing function $\varphi : \mathbb{R} \to [0,\infty)$ has a global minimum. Moreover, let P be a distribution on $X \times Y$ such that $\mathrm{P}(\{x\}) = 0$ for all $x \in X$. We define*

$$\mathcal{S}_{L,\mathrm{P}} := \mathrm{P}\Big(\big\{(x,y) \in X \times Y : 0 \notin \partial\varphi(y\mathcal{M}_{L,\eta(x)}(0^+))\big\}\Big)\,.$$

Furthermore, let k be a bounded measurable and strictly positive definite kernel on X whose RKHS H is separable and satisfies $\mathcal{R}^*_{L,\mathrm{P},H} = \mathcal{R}^*_{L,\mathrm{P}}$. We define $h(\lambda) := \lambda^{-1}|L|_{\lambda^{-1/2},1}$ for $\lambda > 0$. Then, for all $\varepsilon > 0$, there exists a $\delta > 0$ and a $\lambda_0 \in (0,1]$ such that for all $\lambda \in (0,\lambda_0]$ and all $n \geq 1$ the corresponding SVM satisfies

$$\mathrm{P}^n\Big(D \in (X \times Y)^n : \#SV(f_{\mathrm{D},\lambda}) \geq (\mathcal{S}_{L,\mathrm{P}} - \varepsilon)n\Big) \;\geq\; 1 - 2e^{-\frac{\delta^2 n}{18h^2(\lambda)}}\,. \quad (8.31)$$

Note that, roughly speaking, $\mathcal{S}_{L,\mathrm{P}}$ describes the probability of samples (x,y) for which for no Bayes decision function $f^*_{L,\mathrm{P}}$ is the value $yf^*_{L,\mathrm{P}}(x)$ a minimizer of the representing function φ. For example, for the squared hinge loss L, we saw in Exercise 3.1 that

$$\mathcal{M}_{L,\eta}(0^+) = \begin{cases} (-\infty,-1] & \text{if } \eta = 0 \\ \{2\eta - 1\} & \text{if } 0 < \eta < 1 \\ [1,\infty) & \text{if } \eta = 1\,. \end{cases}$$

Moreover, we clearly have $\{t \in \mathbb{R} : 0 \notin \partial\varphi(t)\} = (-\infty,1)$, and hence we obtain $0 \notin \partial\varphi(\pm\mathcal{M}_{L,\eta}(0^+))$ for all $\eta \in (0,1)$. Moreover, for $\eta \in \{0,1\}$, we have $0 \in \partial\varphi(\pm\mathcal{M}_{L,\eta}(0^+))$, and thus we find

$$\mathcal{S}_{L,\mathrm{P}} = \mathrm{P}_X\big(\{x \in X : 0 < \eta(x) < 1\}\big)\,. \quad (8.32)$$

Interestingly, we will see in Theorem 8.36 that for *differentiable* φ the right-hand side of (8.32) is always a lower bound on $\mathcal{S}_{L,\mathrm{P}}$.

Proof. Without loss of generality, we may assume $\varphi(0) \leq 1$ and $\|k\|_\infty \leq 1$. Obviously, $\lambda \mapsto h(\lambda)$ is a decreasing function on $(0,\infty)$, and hence we have $h(\lambda) \geq h(1)$ for all $\lambda \in (0,1]$. Furthermore, note that $\partial L(y,t) = y\partial\varphi(yt)$ yields $\mathcal{S}_{L,\mathrm{P}} = \mathrm{P}(S_0)$, where S_0 is defined in Lemma 8.32. Let us fix an $\varepsilon \in (0,1]$. By

Lemma 8.32, there then exists a $\delta > 0$ such that $\delta \leq 3h(1)\varepsilon$ and $\mathrm{P}(S_{2\delta}) \geq \mathrm{P}(S_0) - \varepsilon$. Moreover, $\lim_{\lambda\to 0} \mathcal{R}_{L,\mathrm{P}}(f_{\mathrm{P},\lambda}) = \mathcal{R}^*_{L,\mathrm{P}}$ together with Theorem 3.61 shows that there exists a $\lambda_0 \in (0,1]$ such that $\mathrm{P}(A(f_{\mathrm{P},\lambda},\delta)) \leq \varepsilon$ for all $\lambda \in (0,\lambda_0]$. Let us fix a $\lambda \in (0,\lambda_0]$ and an $n \geq 1$. Without loss of generality, we may additionally assume

$$\frac{n}{h^2(\lambda)} \geq \frac{18}{\delta^2} \tag{8.33}$$

since otherwise the left-hand side of (8.31) is negative. For $D \in (X \times Y)^n$, we further write

$$\mathcal{E}_{D,\delta} := \bigl|\bigl\{i : (x_i,y_i) \in S_{2\delta}\backslash A(f_{\mathrm{P},\lambda},\delta)\bigr\}\bigr| .$$

Hoeffding's inequality and $\mathrm{P}(S_{2\delta}\backslash A(f_{\mathrm{P},\lambda},\delta)) \geq \mathcal{S}_{L,\mathrm{P}} - 2\varepsilon$ then yield

$$\mathrm{P}^n\Bigl(D \in (X\times Y)^n : \mathcal{E}_{D,\delta} > (\mathcal{S}_{L,\mathrm{P}} - 3\varepsilon)n\Bigr) \;\geq\; 1 - e^{-2\varepsilon^2 n} \;\geq\; 1 - e^{-\frac{\delta^2 n}{18h^2(\lambda)}} ,$$

where in the last step we used that $\delta \leq 3h(1)\varepsilon$ implies $\delta \leq 6h(\lambda)\varepsilon$. Moreover, as in the proof of Theorem 6.24, we have

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

for all $\tau > 0$. We define τ by $\delta = 3h(\lambda)\sqrt{2\tau/n}$. Then $\delta \leq 3h(1)$ together with $h(1) \leq h(\lambda)$ implies $2\tau/n \leq 1$, which in turn yields $\frac{4\tau}{3n} \leq \sqrt{2\tau/n}$. In addition, (8.33) implies $\tau \geq 1$, and hence we have $\sqrt{1/n} \leq \sqrt{2\tau/n}$. Combining these estimates, we conclude that

$$\mathrm{P}^n\bigl(D \in (X\times Y)^n : \|f_{\mathrm{D},\lambda} - f_{\mathrm{P},\lambda}\|_H < \delta\bigr) \geq 1 - e^{-\frac{\delta^2 n}{18h^2(\lambda)}} ,$$

and hence we have

$$\mathrm{P}^n\bigl(D : \|f_{\mathrm{D},\lambda} - f_{\mathrm{P},\lambda}\|_H < \delta \text{ and } \mathcal{E}_{D,\delta} > (\mathcal{S}_{L,\mathrm{P}} - 3\varepsilon)n\bigr) \geq 1 - 2e^{-\frac{\delta^2 n}{18h^2(\lambda)}} .$$

Let us now fix such a $D \in (X \times Y)^n$. For an index i such that $(x_i,y_i) \in S_{2\delta}\backslash A(f_{\mathrm{P},\lambda},\delta)$, Lemma 8.33 then shows that $0 \notin \partial\varphi(y_i f_{\mathrm{D},\lambda}(x_i))$, and hence x_i must be a support vector. Moreover, the estimate above shows that there are at least $(\mathcal{S}_{L,\mathrm{P}} - 3\varepsilon)n$ such indexes. □

In order to illustrate the preceding result assume for simplicity that we use an *a priori* fixed sequence (λ_n) of regularization parameters. If this sequence satisfies both $\lambda_n \to 0$ and $h^2(\lambda_n)/n \to 0$, then the right-hand side of (8.31) converges to 1 and hence the fraction of support vectors tends to be not smaller than $\mathcal{S}_{L,\mathrm{P}}$. Here we note that for the hinge loss, for example, the latter condition on (λ_n) reduces to $\lambda_n^2 n \to \infty$, whereas for the (truncated) least squares loss the condition becomes $\lambda_n^3 n \to \infty$. Finally, it is obvious that

the same result holds for data-dependent choices of λ that asymptotically respect the constraints imposed on (λ_n).

The following proposition presents a general lower bound on $\mathcal{S}_{L,\mathrm{P}}$. Together with Theorem 8.34, it shows that the sparseness of general SVMs is related to the Bayes classification risk.

Proposition 8.35. *Let L be a convex, classification calibrated, and margin-based loss whose representing function $\varphi : \mathbb{R} \to [0,\infty)$ has a global minimum. Then, for all distributions P with $\mathrm{P}_X(\{x \in X : \eta(x) = 1/2\}) = 0$, we have*

$$\mathcal{S}_{L,\mathrm{P}} \;\geq\; \mathcal{R}^*_{L_{\mathrm{class}},\mathrm{P}}\,.$$

Proof. We first show that $0 \notin \partial\varphi(\mathcal{M}_{L,\eta}(0^+)) \cap \partial\varphi(-\mathcal{M}_{L,\eta}(0^+))$ for all $\eta \neq 1/2$. To this end, let us fix an $\eta \in [0,1]$ such that $0 \in \partial\varphi(\mathcal{M}_{L,\eta}(0^+)) \cap \partial\varphi(-\mathcal{M}_{L,\eta}(0^+))$. Then there exist $t, t' \in \mathcal{M}_{L,\eta}(0^+)$ such that $0 \in \partial\varphi(t)$ and $0 \in \partial\varphi(-t')$. From part *ii)* of Lemma 8.31, we conclude that $t' < 0 < t$ and hence we find $H(\eta) = 0$, where $H(\eta)$ is defined in (3.37). From this we conclude that $\eta = 1/2$ by part *i)* of Lemma 3.33. Now the assertion follows from

$$\begin{aligned}\mathcal{S}_{L,\mathrm{P}} &= \mathrm{P}\Big(\big\{(x,y) \in X \times Y : 0 \notin \partial\varphi(y\mathcal{M}_{L,\eta(x)}(0^+))\big\}\Big)\\ &\geq \int_{\eta\neq 1/2} \min\{\eta, 1-\eta\}\, d\mathrm{P}_X\end{aligned}$$

and the formula for $\mathcal{R}^*_{L_{\mathrm{class}},\mathrm{P}}$ given in Example 2.4. □

It is relatively straightforward to show that Proposition 8.35 is sharp for the hinge loss. Combining this with (8.25), we conclude that in general Theorem 8.34 is *not* sharp.

Our next goal is to illustrate the connection between certain desirable properties of L and the asymptotic sparseness of the resulting SVM decision functions. Our first result in this direction shows that differentiable losses L do *not*, in general, lead to sparse decision functions.

Theorem 8.36 (No sparseness for differentiable losses). *Let L be a convex, classification calibrated, and margin-based loss whose representing function $\varphi : \mathbb{R} \to [0,\infty)$ has a global minimum. If φ is differentiable, then for all distributions P we have*

$$\mathcal{S}_{L,\mathrm{P}} \;\geq\; \mathrm{P}_X\big(\{x \in X : 0 < \eta(x) < 1\}\big)\,.$$

Proof. In order to prove the assertion, it obviously suffices to show

$$0 \;\notin\; \partial\varphi\big(\mathcal{M}_{L,\eta}(0^+)\big) \cup \partial\varphi\big(-\mathcal{M}_{L,\eta}(0^+)\big)\,, \qquad \eta \in (0,1).$$

Let us assume the converse, i.e., that there exists an $\eta \in (0,1)$, a $y \in Y$, and a $t \in \mathcal{M}_{L,\eta}(0^+)$ such that $0 \in \partial\varphi(yt)$. Without loss of generality, we may assume $y = 1$. Since L is differentiable, we thus have $\partial L(1,t) = \{0\}$. Therefore, $0 \in \partial\mathcal{C}_{L,\eta}(t)$ implies $0 \in \partial L(-1,t)$, which contradicts part *iii)* of Lemma 8.31. □

Note that certain fast training algorithms (see, e.g., Chapelle, 2007, and the references therein) require the differentiability of the loss. Unfortunately, however, the preceding theorem in conjunction with Theorem 8.34 shows that this possible advantage is paid for by non-sparse decision functions, i.e., by a possible disadvantage during the application phase of the decision function obtained.

At the beginning of this section, we mentioned that another reason for using a loss other than the hinge is the desire to estimate the posterior probability. Our next goal is to show that producing such an estimate is in conflict with the sparseness of the decision functions. We begin with two technical lemmas.

Lemma 8.37. *Let L be a convex, classification calibrated, and margin-based loss whose representing function $\varphi : \mathbb{R} \to [0,\infty)$ has a global minimum. We define $t_0 := \inf\{t \in \mathbb{R} : 0 \in \partial\varphi(t)\}$. Then we have $0 < t_0 < \infty$ and $0 \in \partial\varphi(t_0)$.*

Proof. By Theorem 3.36, we have $t_0 > 0$, and since φ has a global minimum, we further find $t_0 < \infty$. In order to show the third assertion, we observe that by the definition of t_0 there exists a sequence $(s_n) \subset [0,\infty)$ such that $s_n \to 0$ and $0 \in \partial\varphi(t_0 + s_n)$ for all $n \geq 1$. Let us fix an $\varepsilon > 0$. By Proposition A.6.14, there then exists a $\delta > 0$ such that $\partial\varphi(t_0 + \delta B_{\mathbb{R}^d}) \subset \partial\varphi(t_0) + \varepsilon B_{\mathbb{R}^d}$ and for this δ there obviously exists an $n_0 \in \mathbb{N}$ such that $|s_n| \leq \delta$ for all $n \geq n_0$. We conclude that $0 \in \partial\varphi(t_0 + s_n) \subset \partial\varphi(t_0) + \varepsilon B_{\mathbb{R}^d}$, and since $\partial\varphi(t_0)$ is compact, we find $0 \in \partial\varphi(t_0)$ by letting $\varepsilon \to 0$. □

Lemma 8.38. *Let L be a convex, classification calibrated, and margin-based loss function whose representing function $\varphi : \mathbb{R} \to [0,\infty)$ has a global minimum. We define t_0 as in Lemma 8.37 and the function $\gamma : \mathbb{R} \to \mathbb{R}$ by*

$$\gamma(t) := \frac{d^+\varphi(-t)}{d^+\varphi(-t) + d^-\varphi(t)}\,, \qquad t \in \mathbb{R}\,,$$

where d^+ and d^- denote the left and the right derivatives defined in front of Lemma A.6.15. Then we have $1/2 \leq \gamma(t) \leq 1$ for all $t \in [0,t_0]$. Moreover, for all $\eta \in [0,1]$ and all $t \in [0,t_0]$ we have

$$\begin{aligned} \eta \geq \gamma(t) \qquad &\Longleftrightarrow \qquad d^-\mathcal{C}_{L,\eta}(t) \leq 0\,, \\ \eta > \gamma(t) \qquad &\Longleftrightarrow \qquad d^-\mathcal{C}_{L,\eta}(t) < 0\,. \end{aligned}$$

Finally, $d^+\mathcal{C}_{L,\eta}(t_0) \geq 0$ for all $\eta \in [0,1]$, and for $\eta \in [0,1)$ we actually have $d^+\mathcal{C}_{L,\eta}(t_0) > 0$.

Proof. Let us fix a $t \in [0,t_0]$. Then we have $-t \leq 0 < t_0$, and hence both $d^+\varphi(-t)$ and $d^-\varphi(t)$ are negative by the definition of t_0 and the monotonicity of subdifferentials. A simple algebraic transformation thus shows that $\eta \geq \gamma(t)$ if and only if $0 \geq \eta d^-\varphi(t) - (1-\eta)d^+\varphi(-t)$. In addition, we have

$$d^-\mathcal{C}_{L,\eta}(t) = \eta d^-\varphi(t) + (1-\eta)d^-\varphi(-\cdot)(t) = \eta d^-\varphi(t) - (1-\eta)d^+\varphi(-t)\,,$$

and hence we obtain the first equivalence. In addition, a straightforward modification of the preceding proof shows that the equivalence also holds with strict inequalities. For the proof of $\gamma(t) \leq 1$, we first observe that both $d^+\varphi(-t)$ and $d^-\varphi(t)$ are strictly negative for $t \in [0, t_0]$, and hence we find $\gamma(t) \leq 1$. Furthermore, we have $d^-\varphi(t) \geq d^+\varphi(-t)$ by the convexity of φ, and since both derivatives are strictly negative we conclude that $\gamma(t) \geq 1/2$ for all $t \in [0, t_0]$. In order to show the last assertion, we observe by Lemma 8.37 that $0 \in \partial\varphi(t_0) = [d^-\varphi(t_0), d^+(t_0)]$, and hence we have $d^-\varphi(-t_0) \leq d^-(t_0) \leq 0 \leq d^+(t_0)$. Since

$$d^+\mathcal{C}_{L,\eta}(t_0) = \eta d^+\varphi(t_0) - (1-\eta)d^-\varphi(-t_0)\,,$$

we then obtain $d^+\mathcal{C}_{L,\eta}(t_0) \geq 0$. Finally, recall that Theorem 3.36 together with $t_0 \geq 0$ shows that $d^-\varphi(-t_0) < 0$, and hence the argument above yields $d^+\mathcal{C}_{L,\eta}(t_0) > 0$ whenever $\eta \neq 1$. □

With the help of these lemmas, we can now establish a theorem that describes the conflict between the goals of having sparse decision functions and estimating the posterior probability.

Theorem 8.39 (Sparseness vs. estimating posterior probabilities). *Let L be a convex, classification calibrated, and margin-based loss whose representing function $\varphi : \mathbb{R} \to [0, \infty)$ has a global minimum. Let t_0 and γ be defined as in Lemma 8.37 and Lemma 8.38, respectively. Then, for all $\eta \in [0, 1]$ and $y := \operatorname{sign}(2\eta - 1)$, we have*

$$\begin{aligned} 1 - \gamma(t_0) < \eta < \gamma(t_0) \quad &\Longleftrightarrow \quad 0 \notin \partial\varphi(y\mathcal{M}_{L,\eta}(0^+))\,, \\ \eta \notin \{0,1\} \cup [1-\gamma(t_0), \gamma(t_0)] \quad &\Longrightarrow \quad y\mathcal{M}_{L,\eta}(0^+) = \{t_0\}\,, \\ \eta \in \{0,1\} \quad &\Longrightarrow \quad t_0 \in y\mathcal{M}_{L,\eta}(0^+)\,. \end{aligned}$$

Moreover, if the restriction $\varphi_{|(-t_0,t_0)}$ is differentiable and strictly convex, then for all $1-\gamma(t_0) < \eta < \gamma(t_0)$ there exists a $t^*_\eta \in (-t_0, t_0)$ with $\{t^*_\eta\} = \mathcal{M}_{L,\eta}(0^+)$ and $\gamma(t^*_\eta) = \eta$. In addition, the restriction

$$\gamma_{|(-t_0,t_0)} : (-t_0, t_0) \to \big(1-\gamma(t_0), \gamma(t_0)\big)$$

is bijective, continuous, and increasing.

Proof. Since $\mathcal{M}_{L,\eta}(0^+) = -\mathcal{M}_{L,1-\eta}(0^+)$, we observe that we may restrict our considerations to $\eta \in [1/2, 1]$ for the proofs of the first three assertions.

Let us begin with the equivalence. For $\eta \geq \gamma(t_0)$, we find $d^-\mathcal{C}_{L,\eta}(t_0) \leq 0 \leq d^+\mathcal{C}_{L,\eta}(t_0)$ by Lemma 8.38. In other words, we have $0 \in \partial\mathcal{C}_{L,\eta}(t_0)$ and hence $t_0 \in \mathcal{M}_{L,\eta}(0^+)$. Combining this with $0 \in \partial(t_0)$ from Lemma 8.37, we conclude that $0 \in \partial\varphi(\mathcal{M}_{L,\eta}(0^+))$. Conversely, if $\eta \in [1/2, \gamma(t_0))$, we find $d^-\mathcal{C}_{L,\eta}(t_0) > 0$ by Lemma 8.38 and hence $0 \notin \partial\mathcal{C}_{L,\eta}(t_0)$, i.e., $t_0 \notin \mathcal{M}_{L,\eta}(0^+)$. On the other

hand, we have just seen that $t_0 \in \mathcal{M}_{L,\gamma(t_0)}(0^+)$, and since $\eta' \mapsto \mathcal{M}_{L,\eta'}(0^+)$ is a monotone operator by Lemma 8.31, we conclude that $\mathcal{M}_{L,\eta}(0^+) \subset (-\infty, t_0)$. By the definition of t_0, we further have $0 \notin \partial\varphi(t)$ for all $t < t_0$, and hence we find $0 \notin \partial\varphi\bigl(\mathcal{M}_{L,\eta}(0^+)\bigr)$, i.e., we have shown the equivalence.

In order to show the next implication, recall that for $\eta \geq \gamma(t_0)$ we have already seen that $t_0 \in \mathcal{M}_{L,\eta}(0^+)$, and hence it remains to show that $\mathcal{M}_{L,\eta}(0^+) \subset \{t_0\}$ whenever $\gamma(t_0) < \eta < 1$. To show the latter, we first observe that $d^-\mathcal{C}_{L,\eta}(t_0) < 0 < d^+\mathcal{C}_{L,\eta}(t_0)$ by Lemma 8.38. Now assume that $\mathcal{M}_{L,\eta}(0^+)$ contains a $t^* \neq t_0$. If $t^* < t_0$, the monotonicity of $t \mapsto \partial\mathcal{C}_{L,\eta}(t)$ implies $s < 0$ for all $s \in \partial\mathcal{C}_{L,\eta}(t^*)$, which in turn contradicts $t^* \in \mathcal{M}_{L,\eta}(0^+)$. Analogously, we see that $t^* > t_0$ is impossible. The third implication is trivial.

In order to show the remaining assertions, we fix an η with $1-\gamma(t_0) < \eta < \gamma(t_0)$. Then we saw in the first part of the proof that $\mathcal{M}_{L,\eta}(0^+) \subset (-t_0, t_0)$. Let us begin by showing that $\mathcal{M}_{L,\eta}(0^+)$ is a singleton. To this end, we fix $t, t' \in \mathcal{M}_{L,\eta}(0^+)$. Since $\varphi_{|(-t_0,t_0)}$ is differentiable, we then have

$$\begin{aligned} \eta\varphi'(t) - (1-\eta)\varphi'(-t) &= 0\,, \\ \eta\varphi'(t') - (1-\eta)\varphi'(-t') &= 0\,, \end{aligned}$$

and by simple algebra we thus find $\eta(\varphi'(t)-\varphi'(t')) = (1-\eta)(\varphi'(-t)-\varphi'(-t'))$. Let us now assume that $t > t'$. Then the strict convexity of $\varphi_{|(-t_0,t_0)}$ shows both $\varphi'(t) > \varphi'(t')$ and $\varphi'(-t) < \varphi'(-t')$, which clearly contradicts the equality above. Consequently, $\mathcal{M}_{L,\eta}(0^+)$ is indeed a singleton. Now $0 \in \partial\mathcal{C}_{L,\eta}(t^*_\eta)$ together with the definition of γ, the differentiability of $\varphi_{|(-t_0,t_0)}$, and some simple transformations yields $\gamma(t^*_\eta) = \eta$, i.e., $\gamma_{|(-t_0,t_0)}$ is surjective. Conversely, assume that we have a $t \in (-t_0, t_0)$ with $\gamma(t) = \eta$. Then the definition of γ together with the differentiability of $\varphi_{|(-t_0,t_0)}$ and some simple transformations yields $0 \in \partial\mathcal{C}_{L,\eta}(t)$, i.e., $t \in \mathcal{M}_{L,\eta}(0^+) = \{t^*_\eta\}$. Consequently, $\gamma_{|(-t_0,t_0)}$ is also injective. Finally, $\gamma_{|(-t_0,t_0)}$ is continuous since convex, differentiable functions are continuously differentiable by Proposition A.6.14, and $\gamma_{|(-t_0,t_0)}$ is increasing since $\eta \mapsto \mathcal{M}_{L,\eta}(0^+) = \{t^*_\eta\}$ is a monotone operator by Lemma 8.31. □

The preceding theorem is quite technical and merely illuminates the situation. Therefore let us finally illustrate its consequences. To this end, we define

$$G(\eta) := \eta\mathbf{1}_{0\notin\partial\varphi(\mathcal{M}_{L,\eta}(0^+))} + (1-\eta)\mathbf{1}_{0\notin\partial\varphi(\mathcal{M}_{L,1-\eta}(0^+))}\,, \qquad \eta \in [0,1].$$

The definition of $\mathcal{S}_{L,\mathrm{P}}$ together with $-\mathcal{M}_{L,\eta}(0^+) = \mathcal{M}_{L,1-\eta}(0^+)$ then yields

$$\begin{aligned} \mathcal{S}_{L,\mathrm{P}} &= \int_X \eta(x)\mathbf{1}_{0\notin\partial\varphi(\mathcal{M}_{L,\eta(x)}(0^+))} + (1-\eta(x))\mathbf{1}_{0\notin\partial\varphi(-\mathcal{M}_{L,\eta(x)}(0^+))}\,d\mathrm{P}_X(x) \\ &= \int_X G(\eta(x))\,d\mathrm{P}_X(x)\,. \end{aligned}$$

Now the equivalence presented in Theorem 8.39 shows that $G(\eta) = 1$ for all $\eta \in (1-\gamma(t_0), \gamma(t_0))$, and consequently we cannot expect the set

$$\bigl\{x \in X : |2\eta(x) - 1| < 2\gamma(t_0) - 1\bigr\} \tag{8.34}$$

to contribute to the sparseness of the decision function. Let us now assume for a moment that $\varphi_{|(-t_0,t_0)}$ is also differentiable and strictly convex. Then the last part of Theorem 8.39 shows that $\gamma(f^*_{L,\mathrm{P}}(x)) = \eta(x)$ for all $x \in X$ with $-t_0 < f^*_{L,\mathrm{P}}(x) < t_0$. By Corollary 3.62 and the continuity of $\gamma_{|(-t_0,t_0)}$, we thus see that we can estimate $\eta(x)$ whenever x is contained in the set in (8.34). In other words, *if we wish to estimate the posterior probability in regions with high noise, we cannot expect the decision function to have a sparse representation in those regions.* Conversely, the second and third implications of Theorem 8.39 show that on the complement of the set in (8.34), i.e., on the set

$$\bigl\{x \in X : |2\eta(x) - 1| \geq 2\gamma(t_0) - 1\bigr\}, \tag{8.35}$$

it is impossible to estimate the posterior probability using the self-calibration of L. However, Lemma 8.37 shows $0 \in \partial\varphi(t_0)$, and hence we may at least hope to get sparseness on the set in (8.35). In other words, *in regions with low noise, we cannot estimate the posterior probability if we also wish to have sparse representations in these regions.* Finally, note that we can only estimate the posterior probability in regions where the noise is below a certain level, i.e., *if we wish to estimate posterior probability in regions with low noise, we also have to estimate posterior probability in regions with high noise.* We refer to Exercise 8.8 for a loss that only estimates η for noise below a certain level.

8.6 Further Reading and Advanced Topics

Binary classification is one of the oldest problems in machine learning, and consequently there exist a variety of different approaches. A rigorous mathematical treatment of many important methods developed up to the mid 1990s can be found in the book by Devroye *et al.* (1996). In particular, classical methods such as nearest neighbor, histogram rules, and kernel rules are considered in great detail. Moreover, aspects of both classical algorithms and new developments on classification are contained in almost every book on machine learning, so we only mention the recent books by Hastie *et al.* (2001) and Bishop (2006), which give a broad overview of different techniques. Finally, a thorough survey on recent developments on the statistical analysis of classification methods was compiled by Boucheron *et al.* (2005).

The bound on the approximation error function for Gaussian kernels was shown by Steinwart and Scovel (2007) for a slightly more complicated version of the margin-noise exponent. Since at first glance, their concept appears to be closely tailored to the Gaussian kernel, we decided to revise their work. Moreover, it appears that the margin or margin-noise exponent is a somewhat natural concept when dealing with the approximation error function for the hinge loss and *continuous* kernels. Indeed, if we have some noise around the decision boundary, then the minimizer of the hinge loss that we have to

approximate is a step function. Since such functions cannot be uniformly approximated by continuous functions, we have to make some error. Intuitively, this error is larger the closer one gets to the decision boundary. Clearly a low concentration of P_X (and a high noise if (8.13) is used to estimate the excess hinge risk) in this region helps to control the overall error. Finally, Vert and Vert (2006) presented another condition on P that makes it possible to bound the approximation error function of Gaussian kernels.

The notion of the noise exponent goes back to Mammen and Tsybakov (1999) and Tsybakov (2004). Its relation to (a slightly more complicated version of) the margin-noise exponent was described by Steinwart and Scovel (2007). The latter authors also proved the variance bound for the hinge loss; however, the first results in this direction had been shown by Blanchard *et al.* (2008), Massart and Nédélec (2006), and Tarigan and van de Geer (2006). Finally, the resulting improved learning rates for the TV-SVM were found by Steinwart *et al.* (2007) though their results required a substantially finer grid.

It is presently unknown whether the rates obtained are optimal in a minmax sense, i.e., whether there cannot exist a classifier that learns faster than, for example, (8.18) for *all* distributions on $[0,1]^d \times Y$ having a fixed margin-noise exponent β and a fixed noise exponent q. However, learning rates that are faster than those presented are possible under more restrictive assumptions on the data-generating distribution. For example, Steinwart (2001) established *exponentially* fast learning rates for SVMs using either the hinge or the truncated least squares loss if the classes have strictly positive distance from each other and the classification risk is zero. Koltchinskii and Beznosova (2005) generalized this result to distributions having noise exponent $q = \infty$ and a Lipschitz continuous version of the posterior probability. Finally, Audibert and Tsybakov (2007) showed that under certain circumstances such fast rates are also possible for so-called *plug-in* rules.

As in the previous chapters, the parameter selection method discussed is only meant to be an illustration of how the tools work together. We refer to Sections 6.6 and 11.3, where other methods are discussed and references to the literature are given.

One can show that the lower bound (8.25) on the number of support vectors is sometimes sharp for SVMs that use the hinge loss. Namely, Steinwart (2004) proved that, using a Gaussian kernel with *fixed* width γ, there exists a sequence (λ_n) of regularization parameters such that

$$\lim_{n\to\infty} \mathrm{P}^n\left(D \in (X \times Y)^n : \left| \frac{|\#SV(f_{\mathrm{D},\lambda_n})}{n} - 2\mathcal{R}^*_{L_{\mathrm{class}},\mathrm{P}} \right| < \varepsilon \right) = 1$$

for all $\varepsilon > 0$ and all distributions P on $B_{\ell_2^d} \times Y$ whose marginal distributions P_X are absolutely continuous with respect to the Lebesgue measure. However, no particular properties of this sequence were specified, and it is unclear whether this result remains true for certain data-dependent choices of λ (and γ). Steinwart (2004) further showed that the lower bound (8.31) is sharp for

the truncated least squares loss and a fixed Gaussian kernel in the sense that there exists a sequence (λ_n) of regularization parameters such that

$$\lim_{n\to\infty} \mathrm{P}^n\left(D \in (X \times Y)^n : \left|\frac{\#SV(f_{\mathrm{D},\lambda_n})}{n} - \mathcal{S}_{L,P}\right| < \varepsilon\right) = 1\,, \qquad \varepsilon > 0.$$

Here again, P_X is assumed to be absolutely continuous with respect to the Lebesgue measure. Finally, Steinwart (2004) showed in the same sense that $\frac{\#SV(f_{\mathrm{D},\lambda_n})}{n}$ may converge to 1 for the least squares loss. These considerations were extended by Bartlett and Tewari (2004, 2007) to convex, margin-based, classification calibrated losses of the form $\varphi(t) := h((t_0 - t)_+)$, where h is assumed to be continuously differentiable and convex. Namely, they showed under the above-mentioned assumptions of Steinwart (2004) that

$$\lim_{n\to\infty} \mathrm{P}^n\left(D \in (X \times Y)^n : \left|\frac{\#SV(f_{\mathrm{D},\lambda_n})}{n} - \mathbb{E}_{x\sim \mathrm{P}_X} G(\eta(x))\right| < \varepsilon\right) = 1\,,$$

where $G(\eta) := \min\{\eta, 1-\eta\}(1-\eta_0)^{-1}\mathbf{1}_{[0,1-\eta_0]\cup[\eta_0,1]} + \mathbf{1}_{(1-\eta_0,\eta_0)}$ and $\eta_0 := h'(2t_0)/(h'(0)+h'(2t_0))$. Finally, in the presentation of Section 8.4, we followed Steinwart (2004), whereas Section 8.5 is a simplified compilation from both Steinwart (2003) and Bartlett and Tewari (2004, 2007).

There have been some attempts to make SVM decision functions sparser. For example, Wu *et al.* (2005) added a constraint to the primal problem that enforces sparseness. Since their resulting optimization problem is no longer convex, the algorithmic approach is, however, more complicated. A different approach is taken by Bakır *et al.* (2005), who edit the training set in order to remove the samples that are identified as mislabeled. This idea is based on the observation that mislabeled samples will (hopefully) be considered as such by the SVM, and in this case they will be support vectors. Finally, Keerthi *et al.* (2006) consider an SVM that uses the squared hinge loss. To force the decision function to be sparse, they propose to optimize the primal optimization problem over a subset of samples while greedily adding samples to this subset. Finally, references to further approaches can be found in these articles.

The density level detection (DLD) scenario introduced in Example 2.9 can be used for anomaly detection purposes when no labeled data are given. We saw in Section 3.8 that despite these missing labels the DLD learning scenario can be viewed as a binary classification problem and that the classification loss can be used as a surrogate loss. Since, for example, the hinge loss is in turn a surrogate for the classification loss, we can then directly use the hinge loss as a surrogate for the DLD loss function. The details of this approach have been worked out by Steinwart *et al.* (2005) and Scovel *et al.* (2005). A different approach to this problem is taken by the so-called *one-class SVM* proposed by Schölkopf *et al.* (2001a). Remarkably, Vert and Vert (2006) established both consistency and learning rates if this method uses Gaussian kernels with widths depending on the sample size. A learning scenario closely

related to the DLD problem is that of learning minimal volume sets, which was recently investigated by Scott and Nowak (2006). For further work on these learning problems, we refer to the references in the above-mentioned articles.

8.7 Summary

In this chapter, we applied the theory we developed in the previous chapters to analyze the learning performance of SVMs for binary classification. We mainly focused on SVMs using the hinge loss and a Gaussian kernel whose width can change with either the data set size or the data set itself. The key element in this analysis was the margin or margin-noise exponent that described the behavior of the data-generating distribution near the decision boundary. In particular, we saw that, for distributions that have a low concentration and a high noise level in the vicinity of the decision boundary, the approximation error function for Gaussian kernels was relatively small, which in turn resulted in favorable learning rates. We then analyzed a simple strategy for selecting both the regularization parameter and the kernel parameter in a data-dependent way. As in the previous two chapters, it turned out that this strategy is adaptive in the sense that without knowledge about the distribution P it achieves the fastest learning rates that the underlying oracle inequalities can provide. We then improved our analysis by introducing the noise exponent that measures the amount of high noise in the labeling process. This noise exponent implied a variance bound for the hinge loss that in turn could be plugged into our advanced statistical analysis of Chapter 7. The resulting learning rates for the data-dependent parameter selection strategy were sometimes as fast as n^{-1}; however, the exact rates depended heavily on the properties of P.

We then analyzed the typical number of support vectors SVM decision functions have. Here it turned out that, using the hinge loss, the fraction of support vectors tends to be lower bounded by twice the Bayes classification risk. We concluded that we cannot expect extremely sparse decision functions when the underlying distribution has a large Bayes risk.

Since in some situations the hinge loss does not fit the needs of the application, we finally considered alternative surrogate loss functions. Here it turned out that convex margin-based surrogates that do not have a global minimum never lead to sparse decision functions. A typical example of this class of losses is the logistic loss for classification. Moreover, these loss functions are not clippable either, and hence the advanced statistical analysis of Chapter 7 does not apply. For margin-based losses that do have a global minimum, the message was more complicated. First, they do enjoy the advanced statistical analysis of Chapter 7; and second, the established lower bound for the fraction of support vectors is in general smaller than 1. However, we saw examples where this lower bound is not sharp, and hence these results only

suggest that there is a chance of obtaining sparse decision functions. Moreover, for differentiable losses, which are used in some algorithmic approaches, the lower bound actually equals 1 under relatively realistic assumptions on the distribution. Consequently, possible algorithmic advances in the training phase are paid for by disadvantages in the employment phase of the decision function. Finally, we showed that noise levels that may contribute to the sparseness of the decision function cannot be estimated by the decision function. In other words, the ability of estimating the posterior probability by the decision function is paid for by higher costs for the evaluation of the decision function.

8.8 Exercises

8.1. An intuitive illustration for the decision boundary ($\star$)
Let (X, d) be a metric space and P be a distribution on $X \times Y$ that has a *continuous* version of its posterior probability. Show that the corresponding classes X_{-1} and X_1 are open. Furthermore, give some examples where Δ coincides with the intuitive notion of the distance to the decision boundary.

8.2. Checkerboard distributions ($\star$)
Let $m \geq 2$ be a fixed integer and P be the distribution on $[0,1]^d \times Y$ whose marginal distribution P_X is the uniform distribution. Assume that there exists a version of the posterior probability such that every $x = (x_1, \ldots, x_d) \in [0,1]^d$ belongs to the class X_j, where

$$j := \prod_{i=1}^{d} (-1)^{\lceil m x_i \rceil}$$

and $\lceil t \rceil := \min\{n \in \mathbb{N} : n \geq t\}$, $t \in [0, \infty)$. Show that P has margin exponent $\alpha := 1$. Modify P_X so that P has margin exponent α for all $\alpha > 0$.

8.3. Margin exponents and different shapes ($\star$)
For fixed $p > 0$, define $X := \{(x_1, x_2) \in [-1,1]^2 : |x_2| \leq |x_1|^p\}$. Moreover, let P be a distribution on $X \times Y$ whose marginal distribution P_X is the uniform distribution on X. Assume that there exists a version η of the posterior probability such that $\eta(x) > 1/2$ if $x_1 > 0$ and $\eta(x) < 1/2$ if $x_1 < 0$. Draw a picture of the classes X_{-1} and X_1 and show that P has margin exponent $\alpha = 1 + p$.

8.4. Effective classes ($\star\star$)
Show that the margin exponent does not change if we consider the *effective classes* $X_{-1} \cap \operatorname{supp} \mathrm{P}_X$ and $X_1 \cap \operatorname{supp} \mathrm{P}_X$ in the definition of the distance to the decision boundary.

8.5. Another relation between margin and noise exponent ($\star$)
Let (X, d) be a metric space and P be a distribution on $X \times Y$ that has margin exponent $\alpha \in [0, \infty)$. Furthermore, assume that there exist constants $c > 0$ and $\gamma \in [0, \infty)$ and a version η of the posterior probability such that the associated distance to the decision boundary satisfies

$$\Delta(x) \;\le\; c\,|2\eta(x) - 1|^{\gamma}$$

for P_X-almost all $x \in X$. Show that P has noise exponent $q := \alpha\gamma$.

8.6. Sparse decision functions for finite-dimensional RKHSs ($\star$)
Let H be a finite-dimensional RKHS. Show that there always exists a representation (8.19) such that $|\{i : \alpha_i \neq 0\}| \le \dim H$.

8.7. Sparsity for TV-SVMs ($\star\star\star$)
Let X be a compact metric space, L be the hinge loss, H be the RKHS of a universal kernel k over X with $\|k\|_\infty \le 1$, and $a \ge 1$ and $p \in (0, 1]$ be constants with

$$e_i\bigl(\mathrm{id} : H \to C(X)\bigr) \;\le\; a\, i^{-\frac{1}{2p}}\,, \qquad\qquad i \ge 1.$$

Moreover, let $(\varepsilon_n) \subset (0, 1)$ be a sequence with $\varepsilon_n \to 0$ and $\varepsilon_n^{1+p} n \to \infty$. In addition, let $\Lambda = (\Lambda_n)$ be a sequence of ε_n-nets having cardinality growing polynomially in n.

i) Show that the corresponding TV-SVM satisfies the lower bound (8.25) on the number of support vectors.
ii) Generalize this result to TV-SVM using Gaussian kernels with data-dependent kernel parameters.

8.8. Partially estimating the posterior probability ($\star\star\star$)
For a fixed $a \in [0, 1)$, define $\varphi_a : \mathbb{R} \to [0, \infty)$ by $\varphi_a(t) := (1 - t)^2 - a^2$ if $t \le 1 - a$ and $\varphi_a(t) := 0$ otherwise. Show that the margin-based loss represented by φ_a can be used to estimate the posterior probability for noise levels $|2\eta - 1| < 1 - a$. Compare your findings with Theorem 8.39 and discuss the possibility of sparseness.