4 Support Vector Machines

This chapter presents one of the most theoretically well motivated and practically most effective classification algorithms in modern machine learning: Support Vector Machines (SVMs). We first introduce the algorithm for separable datasets, then present its general version designed for non-separable datasets, and finally provide a theoretical foundation for SVMs based on the notion of margin. We start with the description of the problem of linear classification.

4.1 Linear classification

Consider an input space $\mathcal{X}$ that is a subset of $\mathbb{R}^N$ with $N \geq 1$, and the output or target space $\mathcal{Y} = \{-1, +1\}$, and let $f\colon \mathcal{X} \to \mathcal{Y}$ be the target function. Given a hypothesis set H of functions mapping $\mathcal{X}$ to $\mathcal{Y}$, the binary classification task is formulated as follows. The learner receives a training sample S of size m drawn i.i.d. from $\mathcal{X}$ according to some unknown distribution D, $S = ((x_1, y_1), \ldots, (x_m, y_m)) \in (\mathcal{X} \times \mathcal{Y})^m$, with $y_i = f(x_i)$ for all $i \in [1, m]$. The problem consists of determining a hypothesis $h \in H$, a *binary classifier*, with small generalization error:

$$R_D(h) = \Pr_{x \sim D}[h(x) \neq f(x)]. \tag{4.1}$$

Different hypothesis sets H can be selected for this task. In view of the results presented in the previous section, which formalized Occam's razor principle, hypothesis sets with smaller complexity — e.g., smaller VC-dimension or Rademacher complexity — provide better learning guarantees, everything else being equal. A natural hypothesis set with relatively small complexity is that of *linear classifiers*, or hyperplanes, which can be defined as follows:

$$H = \{\mathbf{x} \mapsto \operatorname{sign}(\mathbf{w} \cdot \mathbf{x} + b)\colon \mathbf{w} \in \mathbb{R}^N, b \in \mathbb{R}\}. \tag{4.2}$$

A hypothesis of the form $\mathbf{x} \mapsto \operatorname{sign}(\mathbf{w} \cdot \mathbf{x} + b)$ thus labels positively all points falling on one side of the hyperplane $\mathbf{w} \cdot \mathbf{x} + b = 0$ and negatively all others. The problem is referred to as a *linear classification problem*.

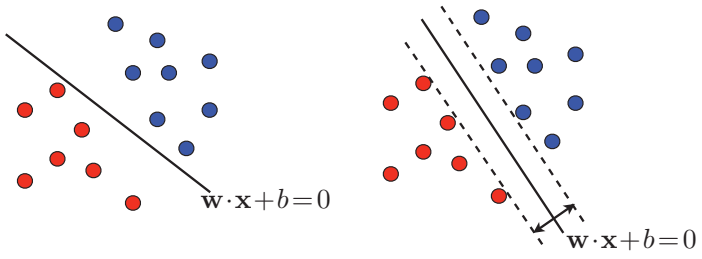

Figure 4.1 Two possible separating hyperplanes. The right-hand side figure shows a hyperplane that maximizes the margin.

4.2 SVMs — separable case

In this section, we assume that the training sample S can be linearly separated, that is, we assume the existence of a hyperplane that perfectly separates the training sample into two populations of positively and negatively labeled points, as illustrated by the left panel of figure 4.1. But there are then infinitely many such separating hyperplanes. Which hyperplane should a learning algorithm select? The solution returned by the SVM algorithm is the hyperplane with the maximum *margin*, or distance to the closest points, and is thus known as the *maximum-margin hyperplane*. The right panel of figure 4.1 illustrates that choice.

We will present later in this chapter a margin theory that provides a strong justification for this solution. We can observe already, however, that the SVM solution can also be viewed as the "safest" choice in the following sense: a test point is classified correctly by a separating hyperplane with margin ρ even when it falls within a distance ρ of the training samples sharing the same label; for the SVM solution, ρ is the maximum margin and thus the "safest" value.

4.2.1 Primal optimization problem

We now derive the equations and optimization problem that define the SVM solution. The general equation of a hyperplane in $\mathbb{R}^N$ is

$$\mathbf{w} \cdot \mathbf{x} + b = 0, \tag{4.3}$$

where $\mathbf{w} \in \mathbb{R}^N$ is a non-zero vector normal to the hyperplane and $b \in \mathbb{R}$ a scalar. Note that this definition of a hyperplane is invariant to non-zero scalar multiplication. Hence, for a hyperplane that does not pass through any sample point, we can scale $\mathbf{w}$ and b appropriately such that $\min_{(\mathbf{x},y)\in S} |\mathbf{w} \cdot \mathbf{x} + b| = 1$.

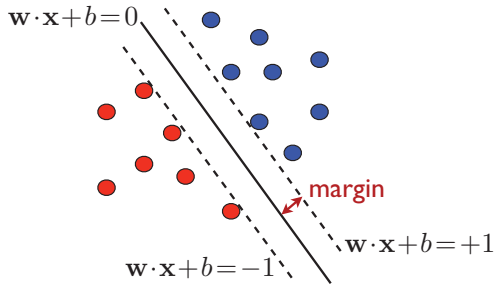

Figure 4.2 Margin and equations of the hyperplanes for a canonical maximum-margin hyperplane. The marginal hyperplanes are represented by dashed lines on the figure.

We define this representation of the hyperplane, i.e., the corresponding pair $(\mathbf{w}, b)$, as the *canonical hyperplane*. The distance of any point $\mathbf{x}_0 \in \mathbb{R}^N$ to a hyperplane defined by (4.3) is given by

$$\frac{|\mathbf{w} \cdot \mathbf{x}_0 + b|}{\|\mathbf{w}\|}. \tag{4.4}$$

Thus, for a canonical hyperplane, the margin ρ is given by

$$\rho = \min_{(\mathbf{x},y)\in S} \frac{|\mathbf{w} \cdot \mathbf{x} + b|}{\|\mathbf{w}\|} = \frac{1}{\|\mathbf{w}\|}. \tag{4.5}$$

Figure 4.2 illustrates the margin for a maximum-margin hyperplane with a canonical representation $(\mathbf{w}, b)$. It also shows the *marginal hyperplanes*, which are the hyperplanes parallel to the separating hyperplane and passing through the closest points on the negative or positive sides. Since they are parallel to the separating hyperplane, they admit the same normal vector $\mathbf{w}$. Furthermore, by definition of a canonical representation, for a point $\mathbf{x}$ on a marginal hyperplane, $|\mathbf{w} \cdot \mathbf{x} + b| = 1$, and thus the equations of the marginal hyperplanes are $\mathbf{w} \cdot \mathbf{x} + b = \pm 1$.

A hyperplane defined by $(\mathbf{w}, b)$ correctly classifies a training point $\mathbf{x}_i$, $i \in [1, m]$ when $\mathbf{w} \cdot \mathbf{x}_i + b$ has the same sign as y_i. For a canonical hyperplane, by definition, we have $|\mathbf{w} \cdot \mathbf{x}_i + b| \geq 1$ for all $i \in [1, m]$; thus, $\mathbf{x}_i$ is correctly classified when $y_i(\mathbf{w} \cdot \mathbf{x}_i + b) \geq 1$. In view of (4.5), maximizing the margin of a canonical hyperplane is equivalent to minimizing $\|\mathbf{w}\|$ or $\frac{1}{2}\|\mathbf{w}\|^2$. Thus, in the separable case, the SVM solution, which is a hyperplane maximizing the margin while correctly classifying all training points, can be expressed as the solution to the following convex optimization problem:

$$\min_{\mathbf{w},b} \frac{1}{2}\|\mathbf{w}\|^2 \tag{4.6}$$
$$\text{subject to: } y_i(\mathbf{w}\cdot\mathbf{x}_i + b) \geq 1,\ \forall i \in [1,m]\,.$$

The objective function $F\colon \mathbf{w} \mapsto \frac{1}{2}\|\mathbf{w}\|^2$ is infinitely differentiable. Its gradient is $\nabla_{\mathbf{w}}(F) = \mathbf{w}$ and its Hessian the identity matrix $\nabla^2 F(\mathbf{w}) = \mathbf{I}$, whose eigenvalues are strictly positive. Therefore, $\nabla^2 F(\mathbf{w}) \succ \mathbf{0}$ and F is strictly convex. The constraints are all defined by affine functions $g_i\colon (\mathbf{w}, b) \mapsto 1 - y_i(\mathbf{w}\cdot\mathbf{x}_i + b)$ and are thus qualified. Thus, in view of the results known for convex optimization (see appendix B for details), the optimization problem of (4.6) admits a unique solution, an important and favorable property that does not hold for all learning algorithms.

Moreover, since the objective function is quadratic and the constraints affine, the optimization problem of (4.6) is in fact a specific instance of *quadratic programming* (QP), a family of problems extensively studied in optimization. A variety of commercial and open-source solvers are available for solving convex QP problems. Additionally, motivated by the empirical success of SVMs along with its rich theoretical underpinnings, specialized methods have been developed to more efficiently solve this particular convex QP problem, notably the block coordinate descent algorithms with blocks of just two coordinates.

4.2.2 Support vectors

The constraints are affine and thus qualified. The objective function as well as the affine constraints are convex and differentiable. Thus, the hypotheses of theorem B.8 hold and the KKT conditions apply at the optimum. We shall use these conditions to both analyze the algorithm and demonstrate several of its crucial properties, and subsequently derive the dual optimization problem associated to SVMs in section 4.2.3.

We introduce Lagrange variables $\alpha_i \geq 0$, $i \in [1,m]$, associated to the m constraints and denote by $\boldsymbol{\alpha}$ the vector $(\alpha_1, \ldots, \alpha_m)^\top$. The Lagrangian can then be defined for all $\mathbf{w} \in \mathbb{R}^N$, $b \in \mathbb{R}$, and $\boldsymbol{\alpha} \in \mathbb{R}^m_+$, by

$$\mathcal{L}(\mathbf{w}, b, \boldsymbol{\alpha}) = \frac{1}{2}\|\mathbf{w}\|^2 - \sum_{i=1}^{m} \alpha_i [y_i(\mathbf{w}\cdot\mathbf{x}_i + b) - 1]\,. \tag{4.7}$$

The KKT conditions are obtained by setting the gradient of the Lagrangian with respect to the primal variables $\mathbf{w}$ and b to zero and by writing the complementarity

conditions:

$$\nabla_{\mathbf{w}}\mathcal{L} = \mathbf{w} - \sum_{i=1}^{m} \alpha_i y_i \mathbf{x}_i = 0 \qquad \Longrightarrow \qquad \mathbf{w} = \sum_{i=1}^{m} \alpha_i y_i \mathbf{x}_i \tag{4.8}$$

$$\nabla_b \mathcal{L} = -\sum_{i=1}^{m} \alpha_i y_i = 0 \qquad \Longrightarrow \qquad \sum_{i=1}^{m} \alpha_i y_i = 0 \tag{4.9}$$

$$\forall i, \alpha_i [y_i(\mathbf{w} \cdot \mathbf{x}_i + b) - 1] = 0 \qquad \Longrightarrow \qquad \alpha_i = 0 \vee y_i(\mathbf{w} \cdot \mathbf{x}_i + b) = 1. \tag{4.10}$$

By equation 4.8, the weight vector $\mathbf{w}$ solution of the SVM problem is a linear combination of the training set vectors $\mathbf{x}_1, \ldots, \mathbf{x}_m$. A vector $\mathbf{x}_i$ appears in that expansion iff $\alpha_i \neq 0$. Such vectors are called *support vectors*. By the complementarity conditions (4.10), if $\alpha_i \neq 0$, then $y_i(\mathbf{w} \cdot \mathbf{x}_i + b) = 1$. Thus, support vectors lie on the marginal hyperplanes $\mathbf{w} \cdot \mathbf{x}_i + b = \pm 1$.

Support vectors fully define the maximum-margin hyperplane or SVM solution, which justifies the name of the algorithm. By definition, vectors not lying on the marginal hyperplanes do not affect the definition of these hyperplanes — in their absence, the solution to the SVM problem remains unchanged. Note that while the solution $\mathbf{w}$ of the SVM problem is unique, the support vectors are not. In dimension N, $N+1$ points are sufficient to define a hyperplane. Thus, when more than $N+1$ points lie on a marginal hyperplane, different choices are possible for the $N+1$ support vectors.

4.2.3 Dual optimization problem

To derive the dual form of the constrained optimization problem (4.6), we plug into the Lagrangian the definition of $\mathbf{w}$ in terms of the dual variables as expressed in (4.8) and apply the constraint (4.9). This yields

$$\mathcal{L} = \underbrace{\frac{1}{2}\|\sum_{i=1}^{m} \alpha_i y_i \mathbf{x}_i\|^2 - \sum_{i,j=1}^{m} \alpha_i \alpha_j y_i y_j (\mathbf{x}_i \cdot \mathbf{x}_j)}_{-\frac{1}{2}\sum_{i,j=1}^{m} \alpha_i \alpha_j y_i y_j (\mathbf{x}_i \cdot \mathbf{x}_j)} \underbrace{- \sum_{i=1}^{m} \alpha_i y_i b}_{0} + \sum_{i=1}^{m} \alpha_i \,, \tag{4.11}$$

which simplifies to

$$\mathcal{L} = \sum_{i=1}^{m} \alpha_i - \frac{1}{2} \sum_{i,j=1}^{m} \alpha_i \alpha_j y_i y_j (\mathbf{x}_i \cdot \mathbf{x}_j) \,. \tag{4.12}$$

This leads to the following dual optimization problem for SVMs in the separable case:

$$\max_{\boldsymbol{\alpha}} \sum_{i=1}^{m} \alpha_i - \frac{1}{2} \sum_{i,j=1}^{m} \alpha_i \alpha_j y_i y_j (\mathbf{x}_i \cdot \mathbf{x}_j) \tag{4.13}$$
$$\text{subject to: } \alpha_i \geq 0 \wedge \sum_{i=1}^{m} \alpha_i y_i = 0, \ \forall i \in [1, m] \,.$$

The objective function $G\colon \boldsymbol{\alpha} \mapsto \sum_{i=1}^{m} \alpha_i - \frac{1}{2} \sum_{i,j=1}^{m} \alpha_i \alpha_j y_i y_j (\mathbf{x}_i \cdot \mathbf{x}_j)$ is infinitely differentiable. Its Hessian is given by $\nabla^2 G = -\mathbf{A}$, with $\mathbf{A} = \big(y_i \mathbf{x}_i \cdot y_j \mathbf{x}_j\big)_{ij}$. $\mathbf{A}$ is the Gram matrix associated to the vectors $y_1 \mathbf{x}_1, \ldots, y_m \mathbf{x}_m$ and is therefore positive semidefinite, which shows that $\nabla^2 G \preceq \mathbf{0}$ and that G is a concave function. Since the constraints are affine and convex, the maximization problem (4.13) is equivalent to a convex optimization problem. Since G is a quadratic function of $\boldsymbol{\alpha}$, this dual optimization problem is also a QP problem, as in the case of the primal optimization and once again both general-purpose and specialized QP solvers can be used to obtain the solution (see exercise 4.4 for details on the SMO algorithm, which is often used to solve the dual form of the SVM problem in the more general non-separable setting).

Moreover, since the constraints are affine, they are qualified and strong duality holds (see appendix B). Thus, the primal and dual problems are equivalent, i.e., the solution $\boldsymbol{\alpha}$ of the dual problem (4.13) can be used directly to determine the hypothesis returned by SVMs, using equation (4.8):

$$h(\mathbf{x}) = \operatorname{sgn}(\mathbf{w} \cdot \mathbf{x} + b) = \operatorname{sgn}\big(\sum_{i=1}^{m} \alpha_i y_i (\mathbf{x}_i \cdot \mathbf{x}) + b\big). \tag{4.14}$$

Since support vectors lie on the marginal hyperplanes, for any support vector $\mathbf{x}_i$, $\mathbf{w} \cdot \mathbf{x}_i + b = y_i$, and thus b can be obtained via

$$b = y_i - \sum_{j=1}^{m} \alpha_j y_j (\mathbf{x}_j \cdot \mathbf{x}_i) \,. \tag{4.15}$$

The dual optimization problem (4.13) and the expressions (4.14) and (4.15) reveal an important property of SVMs: the hypothesis solution depends only on inner products between vectors and not directly on the vectors themselves.

Equation (4.15) can now be used to derive a simple expression of the margin ρ in terms of $\boldsymbol{\alpha}$. Since (4.15) holds for all i with $\alpha_i \neq 0$, multiplying both sides by $\alpha_i y_i$ and taking the sum leads to

$$\sum_{i=1}^{m} \alpha_i y_i b = \sum_{i=1}^{m} \alpha_i y_i^2 - \sum_{i,j=1}^{m} \alpha_i \alpha_j y_i y_j (\mathbf{x}_i \cdot \mathbf{x}_j) \,. \tag{4.16}$$

Using the fact that $y_i^2 = 1$ along with equation 4.8 then yields

$$0 = \sum_{i=1}^{m} \alpha_i - \|\mathbf{w}\|^2. \tag{4.17}$$

Noting that $\alpha_i \geq 0$, we obtain the following expression of the margin ρ in terms of the L_1 norm of $\boldsymbol{\alpha}$:

$$\rho^2 = \frac{1}{\|\mathbf{w}\|_2^2} = \frac{1}{\sum_{i=1}^{m} \alpha_i} = \frac{1}{\|\boldsymbol{\alpha}\|_1}. \tag{4.18}$$

4.2.4 Leave-one-out analysis

We now use the notion of *leave-one-out error* to derive a first learning guarantee for SVMs based on the fraction of support vectors in the training set.

Definition 4.1 **Leave-one-out error**
Let h_S denote the hypothesis returned by a learning algorithm $\mathcal{A}$, when trained on a fixed sample S. Then, the leave-one-out error *of $\mathcal{A}$ on a sample S of size m is defined by*

$$\widehat{R}_{LOO}(\mathcal{A}) = \frac{1}{m} \sum_{i=1}^{m} 1_{h_{S-\{x_i\}}(x_i) \neq y_i}.$$

Thus, for each $i \in [1, m]$, $\mathcal{A}$ is trained on all the points in S except for x_i, i.e., $S - \{x_i\}$, and its error is then computed using x_i. The leave-one-out error is the average of these errors. We will use an important property of the leave-one-out error stated in the following lemma.

Lemma 4.1
The average leave-one-out error for samples of size $m \geq 2$ is an unbiased estimate of the average generalization error for samples of size $m - 1$:

$$\mathop{\mathrm{E}}_{S \sim D^m}[\widehat{R}_{LOO}(\mathcal{A})] = \mathop{\mathrm{E}}_{S' \sim D^{m-1}}[R(h_{S'})], \tag{4.19}$$

where D denotes the distribution according to which points are drawn.

Proof By the linearity of expectation, we can write

$$\begin{aligned}
\mathop{\mathrm{E}}_{S\sim D^m}[\widehat{R}_{\mathrm{LOO}}(\mathcal{A})] &= \frac{1}{m}\sum_{i=1}^{m}\mathop{\mathrm{E}}_{S\sim D^m}[1_{h_{S-\{x_i\}}(x_i)\neq y_i}] \\
&= \mathop{\mathrm{E}}_{S\sim D^m}[1_{h_{S-\{x_1\}}(x_1)\neq y_1}] \\
&= \mathop{\mathrm{E}}_{S'\sim D^{m-1},x_1\sim D}[1_{h_{S'}(x_1)\neq y_1}] \\
&= \mathop{\mathrm{E}}_{S'\sim D^{m-1}}[\mathop{\mathrm{E}}_{x_1\sim D}[1_{h_{S'}(x_1)\neq y_1}]] \\
&= \mathop{\mathrm{E}}_{S'\sim D^{m-1}}[R(h_{S'})].
\end{aligned}$$

For the second equality, we used the fact that, since the points of S are drawn in an i.i.d. fashion, the expectation $\mathrm{E}_{S\sim D^m}[1_{h_{S-\{x_i\}}(x_i)\neq y_i}]$ does not depend on the choice of $i \in [1, m]$ and is thus equal to $\mathrm{E}_{S\sim D^m}[1_{h_{S-\{x_1\}}(x_1)\neq y_1}]$. ■

In general, computing the leave-one-out error may be costly since it requires training m times on samples of size $m-1$. In some situations however, it is possible to derive the expression of $\widehat{R}_{\mathrm{loo}}(\mathcal{A})$ much more efficiently (see exercise 10.9).

Theorem 4.1
Let h_S be the hypothesis returned by SVMs for a sample S, and let $N_{SV}(S)$ be the number of support vectors that define h_S. Then,

$$\mathop{\mathrm{E}}_{S\sim D^m}[R(h_S)] \leq \mathop{\mathrm{E}}_{S\sim D^{m+1}}\left[\frac{N_{\mathrm{SV}}(S)}{m+1}\right].$$

Proof Let S be a linearly separable sample of $m+1$. If x is not a support vector for h_S, removing it does not change the SVM solution. Thus, $h_{S-\{x\}} = h_S$ and $h_{S-\{x\}}$ correctly classifies x. By contraposition, if $h_{S-\{x\}}$ misclassifies x, x must be a support vector, which implies

$$\widehat{R}_{\mathrm{loo}}(\mathrm{SVM}) \leq \frac{N_{\mathrm{SV}}(S)}{m+1}. \tag{4.20}$$

Taking the expectation of both sides and using lemma 4.1 yields the result. ■

Theorem 4.1 gives a sparsity argument in favor of SVMs: the average error of the algorithm is upper bounded by the average fraction of support vectors. One may hope that for many distributions seen in practice, a relatively small number of the training points will lie on the marginal hyperplanes. The solution will then be sparse in the sense that a small fraction of the dual variables α_i will be non-zero. Note, however, that this bound is relatively weak since it applies only to the average generalization error of the algorithm over all samples of size m. It provides no information about the variance of the generalization error. In section 4.4, we present stronger high-probability bounds using a different argument based on the

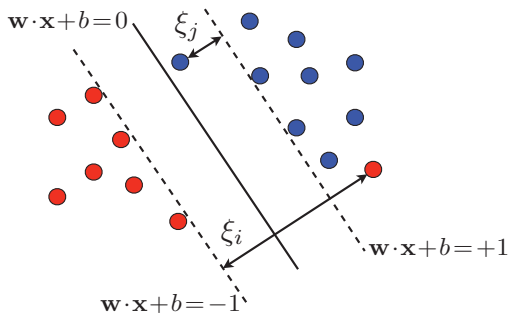

Figure 4.3 A separating hyperplane with point x_i classified incorrectly and point x_j correctly classified, but with margin less than 1.

notion of margin.

4.3 SVMs — non-separable case

In most practical settings, the training data is not linearly separable, i.e., for any hyperplane $\mathbf{w} \cdot \mathbf{x} + b = 0$, there exists $x_i \in S$ such that

$$y_i \left[\mathbf{w} \cdot \mathbf{x}_i + b\right] \ngeq 1 \,. \tag{4.21}$$

Thus, the constraints imposed in the linearly separable case discussed in section 4.2 cannot all hold simultaneously. However, a relaxed version of these constraints can indeed hold, that is, for each $i \in [1, m]$, there exist $\xi_i \geq 0$ such that

$$y_i \left[\mathbf{w} \cdot \mathbf{x}_i + b\right] \geq 1 - \xi_i \,. \tag{4.22}$$

The variables ξ_i are known as *slack variables* and are commonly used in optimization to define relaxed versions of some constraints. Here, a slack variable ξ_i measures the distance by which vector $\mathbf{x}_i$ violates the desired inequality, $y_i(\mathbf{w} \cdot \mathbf{x}_i + b) \geq 1$. Figure 4.3 illustrates the situation. For a hyperplane $\mathbf{w} \cdot \mathbf{x} + b = 0$, a vector $\mathbf{x}_i$ with $\xi_i > 0$ can be viewed as an *outlier*. Each $\mathbf{x}_i$ must be positioned on the correct side of the appropriate marginal hyperplane to not be considered an outlier. As a consequence, a vector $\mathbf{x}_i$ with $0 < y_i(\mathbf{w} \cdot \mathbf{x}_i + b) < 1$ is correctly classified by the hyperplane $\mathbf{w}\cdot\mathbf{x}+b = 0$ but is nonetheless considered to be an outlier, that is, $\xi_i > 0$. If we omit the outliers, the training data is correctly separated by $\mathbf{w} \cdot \mathbf{x} + b = 0$ with a margin $\rho = 1/\|\mathbf{w}\|$ that we refer to as the *soft margin*, as opposed to the *hard margin* in the separable case.

How should we select the hyperplane in the non-separable case? One idea consists of selecting the hyperplane that minimizes the empirical error. But, that solution

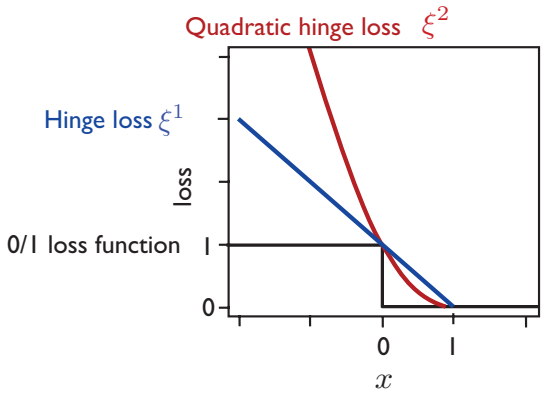

Figure 4.4 Both the hinge loss and the quadratic hinge loss provide convex upper bounds on the binary zero-one loss.

will not benefit from the large-margin guarantees we will present in section 4.4. Furthermore, the problem of determining a hyperplane with the smallest zero-one loss, that is the smallest number of misclassifications, is NP-hard as a function of the dimension N of the space.

Here, there are two conflicting objectives: on one hand, we wish to limit the total amount of slack due to outliers, which can be measured by $\sum_{i=1}^m \xi_i$, or, more generally by $\sum_{i=1}^m \xi_i^p$ for some $p \geq 1$; on the other hand, we seek a hyperplane with a large margin, though a larger margin can lead to more outliers and thus larger amounts of slack.

4.3.1 Primal optimization problem

This leads to the following general optimization problem defining SVMs in the non-separable case where the parameter $C \geq 0$ determines the trade-off between margin-maximization (or minimization of $\|\mathbf{w}\|^2$) and the minimization of the slack penalty $\sum_{i=1}^m \xi_i^p$:

$$\begin{aligned} \min_{\mathbf{w},b,\boldsymbol{\xi}} \quad & \frac{1}{2}\|\mathbf{w}\|^2 + C\sum_{i=1}^m \xi_i^p \\ \text{subject to} \quad & y_i(\mathbf{w}\cdot\mathbf{x}_i + b) \geq 1 - \xi_i \ \wedge \ \xi_i \geq 0, i \in [1,m]\,, \end{aligned} \tag{4.23}$$

where $\boldsymbol{\xi} = (\xi_1, \ldots, \xi_m)^\top$. The parameter C is typically determined via n-fold cross-validation (see section 1.3).

As in the separable case, (4.23) is a convex optimization problem since the constraints are affine and thus convex and since the objective function is convex for any $p \geq 1$. In particular, $\boldsymbol{\xi} \mapsto \sum_{i=1}^m \xi_i^p = \|\boldsymbol{\xi}\|_p^p$ is convex in view of the convexity of the norm $\|\cdot\|_p$.

There are many possible choices for p leading to more or less aggressive penalizations of the slack terms (see exercise 4.1). The choices $p = 1$ and $p = 2$ lead to the most straightforward solutions and analyses. The loss functions associated with $p = 1$ and $p = 2$ are called the *hinge loss* and the *quadratic hinge loss*, respectively. Figure 4.4 shows the plots of these loss functions as well as that of the standard zero-one loss function. Both hinge losses are convex upper bounds on the zero-one loss, thus making them well suited for optimization. In what follows, the analysis is presented in the case of the hinge loss ($p = 1$), which is the most widely used loss function for SVMs.

4.3.2 Support vectors

As in the separable case, the constraints are affine and thus qualified. The objective function as well as the affine constraints are convex and differentiable. Thus, the hypotheses of theorem B.8 hold and the KKT conditions apply at the optimum. We use these conditions to both analyze the algorithm and demonstrate several of its crucial properties, and subsequently derive the dual optimization problem associated to SVMs in section 4.3.3.

We introduce Lagrange variables $\alpha_i \geq 0$, $i \in [1, m]$, associated to the first m constraints and $\beta_i \geq 0$, $i \in [1, m]$ associated to the non-negativity constraints of the slack variables. We denote by $\boldsymbol{\alpha}$ the vector $(\alpha_1, \ldots, \alpha_m)^\top$ and by $\boldsymbol{\beta}$ the vector $(\beta_1, \ldots, \beta_m)^\top$. The Lagrangian can then be defined for all $\mathbf{w} \in \mathbb{R}^N$, $b \in \mathbb{R}$, and $\boldsymbol{\alpha} \in \mathbb{R}^m_+$, by

$$\mathcal{L}(\mathbf{w}, b, \boldsymbol{\xi}, \boldsymbol{\alpha}, \boldsymbol{\beta}) = \frac{1}{2}\|\mathbf{w}\|^2 + C\sum_{i=1}^{m}\xi_i - \sum_{i=1}^{m}\alpha_i[y_i(\mathbf{w}\cdot\mathbf{x}_i + b) - 1 + \xi_i] - \sum_{i=1}^{m}\beta_i\xi_i\,. \quad (4.24)$$

The KKT conditions are obtained by setting the gradient of the Lagrangian with respect to the primal variables $\mathbf{w}$, b, and ξ_is to zero and by writing the complementarity conditions:

$$\nabla_{\mathbf{w}}\mathcal{L} = \mathbf{w} - \sum_{i=1}^{m}\alpha_i y_i \mathbf{x}_i = 0 \quad \Longrightarrow \quad \mathbf{w} = \sum_{i=1}^{m}\alpha_i y_i \mathbf{x}_i \quad (4.25)$$

$$\nabla_b\mathcal{L} = -\sum_{i=1}^{m}\alpha_i y_i = 0 \quad \Longrightarrow \quad \sum_{i=1}^{m}\alpha_i y_i = 0 \quad (4.26)$$

$$\nabla_{\xi_i}\mathcal{L} = C - \alpha_i - \beta_i = 0 \quad \Longrightarrow \quad \alpha_i + \beta_i = C \quad (4.27)$$

$$\forall i, \alpha_i[y_i(\mathbf{w}\cdot\mathbf{x}_i + b) - 1 + \xi_i] = 0 \quad \Longrightarrow \quad \alpha_i = 0 \vee y_i(\mathbf{w}\cdot\mathbf{x}_i + b) = 1 - \xi_i \quad (4.28)$$

$$\forall i, \beta_i\xi_i = 0 \quad \Longrightarrow \quad \beta_i = 0 \vee \xi_i = 0\,. \quad (4.29)$$

By equation 4.25, as in the separable case, the weight vector $\mathbf{w}$ solution of the SVMproblem is a linear combination of the training set vectors $\mathbf{x}_1, \ldots, \mathbf{x}_m$. A vector

$\mathbf{x}_i$ appears in that expansion iff $\alpha_i \neq 0$. Such vectors are called *support vectors*. Here, there are two types of support vectors. By the complementarity condition (4.28), if $\alpha_i \neq 0$, then $y_i(\mathbf{w} \cdot \mathbf{x}_i + b) = 1 - \xi_i$. If $\xi_i = 0$, then $y_i(\mathbf{w} \cdot \mathbf{x}_i + b) = 1$ and $\mathbf{x}_i$ lies on a marginal hyperplane, as in the separable case. Otherwise, $\xi_i \neq 0$ and $\mathbf{x}_i$ is an outlier. In this case, (4.29) implies $\beta_i = 0$ and (4.27) then requires $\alpha_i = C$. Thus, support vectors $\mathbf{x}_i$ are either outliers, in which case $\alpha_i = C$, or vectors lying on the marginal hyperplanes. As in the separable case, note that while the weight vector $\mathbf{w}$ solution is unique, the support vectors are not.

4.3.3 Dual optimization problem

To derive the dual form of the constrained optimization problem (4.23), we plug into the Lagrangian the definition of $\mathbf{w}$ in terms of the dual variables (4.25) and apply the constraint (4.26). This yields

$$\mathcal{L} = \underbrace{\frac{1}{2}\|\sum_{i=1}^{m} \alpha_i y_i \mathbf{x}_i\|^2 - \sum_{i,j=1}^{m} \alpha_i \alpha_j y_i y_j (\mathbf{x}_i \cdot \mathbf{x}_j)}_{-\frac{1}{2}\sum_{i,j=1}^{m} \alpha_i \alpha_j y_i y_j (\mathbf{x}_i \cdot \mathbf{x}_j)} - \underbrace{\sum_{i=1}^{m} \alpha_i y_i b}_{0} + \sum_{i=1}^{m} \alpha_i \,. \tag{4.30}$$

Remarkably, we find that the objective function is no different than in the separable case:

$$\mathcal{L} = \sum_{i=1}^{m} \alpha_i - \frac{1}{2} \sum_{i,j=1}^{m} \alpha_i \alpha_j y_i y_j (\mathbf{x}_i \cdot \mathbf{x}_j) \,. \tag{4.31}$$

However, here, in addition to $\alpha_i \geq 0$, we must impose the constraint on the Lagrange variables $\beta_i \geq 0$. In view of (4.27), this is equivalent to $\alpha_i \leq C$. This leads to the following dual optimization problem for SVMs in the non-separable case, which only differs from that of the separable case (4.13) by the constraints $\alpha_i \leq C$:

$$\max_{\boldsymbol{\alpha}} \sum_{i=1}^{m} \alpha_i - \frac{1}{2} \sum_{i,j=1}^{m} \alpha_i \alpha_j y_i y_j (\mathbf{x}_i \cdot \mathbf{x}_j) \tag{4.32}$$

$$\text{subject to: } 0 \leq \alpha_i \leq C \wedge \sum_{i=1}^{m} \alpha_i y_i = 0, i \in [1, m].$$

Thus, our previous comments about the optimization problem (4.13) apply to (4.32) as well. In particular, the objective function is concave and infinitely differentiable and (4.32) is equivalent to a convex QP. The problem is equivalent to the primal problem (4.23).

The solution $\boldsymbol{\alpha}$ of the dual problem (4.32) can be used directly to determine the

hypothesis returned by SVMs, using equation (4.25):

$$h(\mathbf{x}) = \operatorname{sgn}(\mathbf{w} \cdot \mathbf{x} + b) = \operatorname{sgn}\big(\sum_{i=1}^{m} \alpha_i y_i (\mathbf{x}_i \cdot \mathbf{x}) + b\big). \tag{4.33}$$

Moreover, b can be obtained from any support vector $\mathbf{x}_i$ lying on a marginal hyperplane, that is any vector $\mathbf{x}_i$ with $0 < \alpha_i < C$. For such support vectors, $\mathbf{w} \cdot \mathbf{x}_i + b = y_i$ and thus

$$b = y_i - \sum_{j=1}^{m} \alpha_j y_j (\mathbf{x}_j \cdot \mathbf{x}_i)\,. \tag{4.34}$$

As in the separable case, the dual optimization problem (4.32) and the expressions (4.33) and (4.34) show an important property of SVMs: the hypothesis solution depends only on inner products between vectors and not directly on the vectors themselves. This fact can be used to extend SVMs to define non-linear decision boundaries, as we shall see in chapter 5.

4.4 Margin theory

This section presents generalization bounds based on the notion of margin, which provide a strong theoretical justification for the SVM algorithm. We first give the definitions of some basic margin concepts.

Definition 4.2 **Margin**
The geometric margin $\rho(x)$ *of a point* $\mathbf{x}$ *with label* y *with respect to a linear classifier* $h\colon \mathbf{x} \mapsto \mathbf{w} \cdot \mathbf{x} + b$ *is its distance to the hyperplane* $\mathbf{w} \cdot \mathbf{x} + b = 0$*:*

$$\rho(x) = \frac{y\,(\mathbf{w} \cdot \mathbf{x} + b)}{\|\mathbf{w}\|}\,. \tag{4.35}$$

The margin *of a linear classifier* h *for a sample* $S = (\mathbf{x}_1, \ldots, \mathbf{x}_m)$ *is the minimum margin over the points in the sample:*

$$\rho = \min_{1 \le i \le m} \frac{y_i\,(\mathbf{w} \cdot \mathbf{x}_i + b)}{\|\mathbf{w}\|}\,. \tag{4.36}$$

Recall that the VC-dimension of the family of hyperplanes or linear hypotheses in $\mathbb{R}^N$ is $N+1$. Thus, the application of the VC-dimension bound (3.31) of corollary 3.4 to this hypothesis set yields the following: for any $\delta > 0$, with probability at least

$1-\delta$, for any $h \in H$,

$$R(h) \leq \widehat{R}(h) + \sqrt{\frac{2(N+1)\log\frac{em}{N+1}}{m}} + \sqrt{\frac{\log\frac{1}{\delta}}{2m}}. \tag{4.37}$$

When the dimension of the feature space N is large compared to the sample size, this bound is uninformative. The following theorem presents instead a bound on the VC-dimension of canonical hyperplanes that does not depend on the dimension of feature space N, but only on the margin and the radius r of the sphere containing the data.

Theorem 4.2
Let $S \subseteq \{\mathbf{x}\colon \|\mathbf{x}\| \leq r\}$. Then, the VC-dimension d of the set of canonical hyperplanes $\{x \mapsto \operatorname{sgn}(\mathbf{w}\cdot\mathbf{x})\colon \min_{x\in S}|\mathbf{w}\cdot\mathbf{x}| = 1 \wedge \|\mathbf{w}\| \leq \Lambda\}$ verifies

$$d \leq r^2\Lambda^2\,.$$

Proof Assume $\{\mathbf{x}_1,\ldots,\mathbf{x}_d\}$ is a set that can be fully shattered. Then, for all $\mathbf{y} = (y_1,\ldots,y_d) \in \{-1,+1\}^d$, there exists $\mathbf{w}$ such that,

$$\forall i \in [1,d], 1 \leq y_i(\mathbf{w}\cdot\mathbf{x}_i)\,.$$

Summing up these inequalities yields

$$d \leq \mathbf{w}\cdot\sum_{i=1}^{d} y_i\mathbf{x}_i \leq \|\mathbf{w}\|\|\sum_{i=1}^{d} y_i\mathbf{x}_i\| \leq \Lambda\|\sum_{i=1}^{d} y_i\mathbf{x}_i\|\,.$$

Since this inequality holds for all $\mathbf{y} \in \{-1,+1\}^d$, it also holds on expectation over $y_1,\ldots,y_d$ drawn i.i.d. according to a uniform distribution over $\{-1,+1\}$. In view of the independence assumption, for $i \neq j$ we have $\mathrm{E}[y_iy_j] = \mathrm{E}[y_i]\,\mathrm{E}[y_j]$. Thus, since the distribution is uniform, $\mathrm{E}[y_iy_j] = 0$ if $i \neq j$, $\mathrm{E}[y_iy_j] = 1$ otherwise. This gives

$$\begin{aligned}
d &\leq \Lambda \mathop{\mathrm{E}}_{\mathbf{y}}[\|\sum_{i=1}^{d} y_i\mathbf{x}_i\|] && \text{(taking expectations)}\\
&\leq \Lambda\Big[\mathop{\mathrm{E}}_{\mathbf{y}}[\|\sum_{i=1}^{d} y_i\mathbf{x}_i\|^2]\Big]^{1/2} && \text{(Jensen's inequality)}\\
&= \Lambda\Big[\sum_{i,j=1}^{d} \mathop{\mathrm{E}}_{\mathbf{y}}[y_iy_j](\mathbf{x}_i\cdot\mathbf{x}_j)\Big]^{1/2}\\
&= \Lambda\big[\sum_{i=1}^{d}(\mathbf{x}_i\cdot\mathbf{x}_i)\big]^{1/2} \leq \Lambda\big[dr^2\big]^{1/2} = \Lambda r\sqrt{d}.
\end{aligned}$$

Thus, $\sqrt{d} \leq \Lambda r$, which completes the proof. ■

When the training data is linearly separable, by the results of section 4.2, the maximum-margin canonical hyperplane with $\|\mathbf{w}\| = 1/\rho$ can be plugged into theorem 4.2. In this case, Λ can be set to $1/\rho$, and the upper bound can be rewritten as r^2/ρ^2. Note that the choice of Λ must be made before receiving the sample S.

It is also possible to bound the Rademacher complexity of linear hypotheses with bounded weight vector in a similar way, as shown by the following theorem.

Theorem 4.3
Let $S \subseteq \{x\colon \|\mathbf{x}\| \leq R\}$ be a sample of size m and let $H = \{\mathbf{x} \mapsto \mathbf{w} \cdot \mathbf{x}\colon \|\mathbf{w}\| \leq \Lambda\}$. Then, the empirical Rademacher complexity of H can be bounded as follows:

$$\widehat{\mathfrak{R}}_S(H) \leq \sqrt{\frac{r^2\Lambda^2}{m}}\,.$$

Proof The proof follows through a series of inequalities similar to those of theorem 4.2:

$$\begin{aligned}
\widehat{\mathfrak{R}}_S(H) &= \frac{1}{m}\mathop{\mathrm{E}}_{\sigma}\Big[\sum_{i=1}^{m}\sigma_i\mathbf{w}\cdot\mathbf{x}_i\Big] = \frac{1}{m}\mathop{\mathrm{E}}_{\sigma}\Big[\mathbf{w}\cdot\sum_{i=1}^{m}\sigma_i\mathbf{x}_i\Big] \leq \frac{\Lambda}{m}\mathop{\mathrm{E}}_{\sigma}\Big[\Big\|\sum_{i=1}^{m}\sigma_i\mathbf{x}_i\Big\|\Big] \\
&\leq \frac{\Lambda}{m}\left[\mathop{\mathrm{E}}_{\sigma}\Big[\Big\|\sum_{i=1}^{m}\sigma_i\mathbf{x}_i\Big\|^2\Big]\right]^{1/2} = \frac{\Lambda}{m}\left[\mathop{\mathrm{E}}_{\sigma}\Big[\sum_{i,j=1}^{m}\sigma_i\sigma_j(\mathbf{x}_i\cdot\mathbf{x}_j)\Big]\right]^{1/2} \\
&\leq \frac{\Lambda}{m}\left[\sum_{i=1}^{m}\|\mathbf{x}_i\|^2\right]^{1/2} \leq \frac{\Lambda\sqrt{mr^2}}{m} = \sqrt{\frac{r^2\Lambda^2}{m}}\,,
\end{aligned}$$

The first inequality makes use of the Cauchy-Schwarz inequality and the bound on $\|\mathbf{w}\|$, the second follows by Jensen's inequality, the third by $\mathrm{E}[\sigma_i\sigma_j] = \mathrm{E}[\sigma_i]\,\mathrm{E}[\sigma_j] = 0$ for $i \neq j$, and the last one by $\|\mathbf{x}_i\| \leq R$. ■

To present the main margin-based generalization bounds of this section, we need to introduce a *margin loss function*. Here, the training data is not assumed to be separable. The quantity $\rho > 0$ should thus be interpreted as the margin we wish to achieve.

Definition 4.3 **Margin loss function**
For any $\rho > 0$, the ρ-margin loss is the function $L_\rho\colon \mathbb{R}\times\mathbb{R} \to \mathbb{R}_+$ defined for all $y, y' \in \mathbb{R}$ by $L_\rho(y, y') = \Phi_\rho(yy')$ with,

$$\Phi_\rho(x) = \begin{cases} 0 & \text{if } \rho \leq x \\ 1 - x/\rho & \text{if } 0 \leq x \leq \rho \\ 1 & \text{if } x \leq 0\,. \end{cases}$$

This loss function is illustrated in figure 4.5. The empirical margin loss is then defined as the margin loss over the training sample.

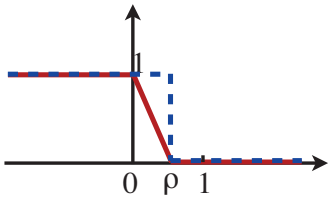

Figure 4.5 The margin loss, defined with respect to margin parameter ρ.

Definition 4.4 **Empirical margin loss**
Given a sample $S = (x_1, \ldots, x_m)$ and a hypothesis h, the empirical margin loss is defined by

$$\widehat{R}_\rho(h) = \frac{1}{m} \sum_{i=1}^{m} \Phi_\rho(y_i h(x_i)) \,. \tag{4.38}$$

Note that for any $i \in [1, m]$, $\Phi_\rho(y_i h(x_i)) \leq 1_{y_i h(x_i) \leq \rho}$. Thus, the empirical margin loss can be upper-bounded as follows:

$$\widehat{R}_\rho(h) \leq \frac{1}{m} \sum_{i=1}^{m} 1_{y_i h(x_i) \leq \rho} \,. \tag{4.39}$$

In all the results that follow, the empirical margin loss can be replaced by this upper bound, which admits a simple interpretation: it is the fraction of the points in the training sample S that have been misclassified or classified with confidence less than ρ. When h is a linear function defined by a weight vector $\mathbf{w}$ with $\|\mathbf{w}\| = 1$, $y_i h(x_i)$ is the margin of point x_i. Thus, the upper bound is then the fraction of the points in the training data with margin less than ρ. This corresponds to the loss function indicated by the blue dotted line in figure 4.5.

The slope of the function Φ_ρ defining the margin loss is at most $1/\rho$, thus Φ_ρ is $1/\rho$-Lipschitz. The following lemma bounds the empirical Rademacher complexity of a hypothesis set H after composition with such a Lipschitz function in terms of the empirical Rademacher complexity of H. It will be needed for the proof of the margin-based generalization bound.

Lemma 4.2 **Talagrand's lemma**
Let $\Phi \colon \mathbb{R} \to \mathbb{R}$ be an l-Lipschitz. Then, for any hypothesis set H of real-valued functions, the following inequality holds:

$$\widehat{\mathfrak{R}}_S(\Phi \circ H) \leq l \, \widehat{\mathfrak{R}}_S(H) \,.$$

Proof First we fix a sample $S = (x_1, \ldots, x_m)$, then, by definition,

$$\begin{aligned}\widehat{\mathfrak{R}}_S(\Phi \circ H) &= \frac{1}{m} \mathop{\mathrm{E}}_{\boldsymbol{\sigma}} \Big[\sup_{h \in H} \sum_{i=1}^{m} \sigma_i (\Phi \circ h)(x_i) \Big] \\ &= \frac{1}{m} \mathop{\mathrm{E}}_{\sigma_1, \ldots, \sigma_{m-1}} \Big[\mathop{\mathrm{E}}_{\sigma_m} \Big[\sup_{h \in H} u_{m-1}(h) + \sigma_m (\Phi \circ h)(x_m) \Big] \Big],\end{aligned}$$

where $u_{m-1}(h) = \sum_{i=1}^{m-1} \sigma_i (\Phi \circ h)(x_i)$. By definition of the supremum, for any $\epsilon > 0$, there exist $h_1, h_2 \in H$ such that

$$\begin{aligned} u_{m-1}(h_1) + (\Phi \circ h_1)(x_m) &\geq (1-\epsilon) \Big[\sup_{h \in H} u_{m-1}(h) + (\Phi \circ h)(x_m) \Big] \\ \text{and} \quad u_{m-1}(h_2) - (\Phi \circ h_2)(x_m) &\geq (1-\epsilon) \Big[\sup_{h \in H} u_{m-1}(h) - (\Phi \circ h)(x_m) \Big]. \end{aligned}$$

Thus, for any $\epsilon > 0$, by definition of E_{σ_m},

$$\begin{aligned}&(1-\epsilon) \mathop{\mathrm{E}}_{\sigma_m} \Big[\sup_{h \in H} u_{m-1}(h) + \sigma_m (\Phi \circ h)(x_m) \Big] \\ &\quad = (1-\epsilon) \Big[\frac{1}{2} \sup_{h \in H} u_{m-1}(h) + (\Phi \circ h)(x_m) + \frac{1}{2} \sup_{h \in H} u_{m-1}(h) - (\Phi \circ h)(x_m) \Big] \\ &\quad \leq \frac{1}{2} [u_{m-1}(h_1) + (\Phi \circ h_1)(x_m)] + \frac{1}{2} [u_{m-1}(h_2) - (\Phi \circ h_2)(x_m)].\end{aligned}$$

Let $s = \operatorname{sgn}(h_1(x_m) - h_2(x_m))$. Then, the previous inequality implies

$$\begin{aligned}&(1-\epsilon) \mathop{\mathrm{E}}_{\sigma_m} \Big[\sup_{h \in H} u_{m-1}(h) + \sigma_m (\Phi \circ h)(x_m) \Big] \\ &\leq \frac{1}{2} [u_{m-1}(h_1) + u_{m-1}(h_2) + sl(h_1(x_m) - h_2(x_m))] && \text{(Lipschitz property)} \\ &= \frac{1}{2} [u_{m-1}(h_1) + slh_1(x_m)] + \frac{1}{2} [u_{m-1}(h_2) - slh_2(x_m)] && \text{(rearranging)} \\ &\leq \frac{1}{2} \sup_{h \in H} [u_{m-1}(h) + slh(x_m)] + \frac{1}{2} \sup_{h \in H} [u_{m-1}(h) - slh(x_m)] && \text{(definition of sup)} \\ &= \mathop{\mathrm{E}}_{\sigma_m} \Big[\sup_{h \in H} u_{m-1}(h) + \sigma_m lh(x_m) \Big]. && \text{(definition of } \mathop{\mathrm{E}}_{\sigma_m} \text{)}\end{aligned}$$

Since the inequality holds for all $\epsilon > 0$, we have

$$\mathop{\mathrm{E}}_{\sigma_m} \Big[\sup_{h \in H} u_{m-1}(h) + \sigma_m (\Phi \circ h)(x_m) \Big] \leq \mathop{\mathrm{E}}_{\sigma_m} \Big[\sup_{h \in H} u_{m-1}(h) + \sigma_m lh(x_m) \Big].$$

Proceeding in the same way for all other σ_is ($i \neq m$) proves the lemma. ■

The following is a general margin-based generalization bound that will be used in the analysis of several algorithms.

Theorem 4.4 **Margin bound for binary classification**
Let H be a set of real-valued functions. Fix $\rho > 0$, then, for any $\delta > 0$, with probability at least $1 - \delta$, each of the following holds for all $h \in H$:

$$R(h) \leq \widehat{R}_\rho(h) + \frac{2}{\rho}\mathfrak{R}_m(H) + \sqrt{\frac{\log \frac{1}{\delta}}{2m}} \tag{4.40}$$

$$R(h) \leq \widehat{R}_\rho(h) + \frac{2}{\rho}\widehat{\mathfrak{R}}_S(H) + 3\sqrt{\frac{\log \frac{2}{\delta}}{2m}}\,. \tag{4.41}$$

Proof Let $\widetilde{H} = \{z = (x, y) \mapsto yh(x) \colon h \in H\}$. Consider the family of functions taking values in $[0, 1]$:

$$\widetilde{\mathcal{H}} = \{\Phi_\rho \circ f \colon f \in \widetilde{H}\}\,.$$

By theorem 3.1, with probability at least $1 - \delta$, for all $g \in \widetilde{\mathcal{H}}$,

$$\mathrm{E}[g(z)] \leq \frac{1}{m}\sum_{i=1}^{m} g(z_i) + 2\mathfrak{R}_m(\widetilde{\mathcal{H}}) + \sqrt{\frac{\log \frac{1}{\delta}}{2m}}\,,$$

and thus, for all $h \in H$,

$$\mathrm{E}[\Phi_\rho(yh(x))] \leq \widehat{R}_\rho(h) + 2\mathfrak{R}_m\big(\Phi_\rho \circ \widetilde{H}\big) + \sqrt{\frac{\log \frac{1}{\delta}}{2m}}\,.$$

Since $1_{u \leq 0} \leq \Phi_\rho(u)$ for all $u \in \mathbb{R}$, we have $R(h) = \mathrm{E}[1_{yh(x) \leq 0}] \leq \mathrm{E}[\Phi_\rho(yh(x))]$, thus

$$R(h) \leq \widehat{R}_\rho(h) + 2\mathfrak{R}_m\big(\Phi_\rho \circ \widetilde{H}\big) + \sqrt{\frac{\log \frac{1}{\delta}}{2m}}\,.$$

$\mathfrak{R}_m$ is invariant to a constant shift, therefore we have

$$\mathfrak{R}_m\big(\Phi_\rho \circ \widetilde{H}\big) = \mathfrak{R}_m\big((\Phi_\rho - 1) \circ \widetilde{H}\big).$$

Since $(\Phi_\rho - 1)(0) = 0$ and since $(\Phi_\rho - 1)$ is $1/\rho$-Lipschitz as with Φ_ρ, by lemma 4.2, we have $\mathfrak{R}_m\big(\Phi_\rho \circ \widetilde{H}\big) \leq \frac{1}{\rho}\mathfrak{R}_m(\widetilde{H})$ and $\mathfrak{R}_m(\widetilde{H})$ can be rewritten as follows:

$$\mathfrak{R}_m(\widetilde{H}) = \frac{1}{m}\mathop{\mathrm{E}}_{S,\sigma}\Big[\sup_{h \in H}\sum_{i=1}^{m}\sigma_i y_i h(x_i)\Big] = \frac{1}{m}\mathop{\mathrm{E}}_{S,\sigma}\Big[\sup_{h \in H}\sum_{i=1}^{m}\sigma_i h(x_i)\Big] = \mathfrak{R}_m\big(H\big)\,.$$

This proves (4.40). The second inequality, (4.41), can be derived in the same way by using the second inequality of theorem 3.1, (3.4), instead of (3.3). ■

The generalization bounds of theorem 4.4 shows the conflict between two terms: the larger the desired margin ρ, the smaller the middle term; however, the first

term, the empirical margin loss $\widehat{R}_\rho$, increases as a function of ρ. The bounds of this theorem can be generalized to hold uniformly for all $\rho > 0$ at the cost of an additional term $\sqrt{\frac{\log\log_2 \frac{2}{\rho}}{m}}$, as shown in the following theorem (a version of this theorem with better constants can be derived, see exercise 4.2).

Theorem 4.5
Let H be a set of real-valued functions. Then, for any $\delta > 0$, with probability at least $1 - \delta$, each of the following holds for all $h \in H$ and $\rho \in (0, 1)$:

$$R(h) \leq \widehat{R}_\rho(h) + \frac{4}{\rho}\mathfrak{R}_m(H) + \sqrt{\frac{\log\log_2 \frac{2}{\rho}}{m}} + \sqrt{\frac{\log \frac{2}{\delta}}{2m}} \tag{4.42}$$

$$R(h) \leq \widehat{R}_\rho(h) + \frac{4}{\rho}\widehat{\mathfrak{R}}_S(H) + \sqrt{\frac{\log\log_2 \frac{2}{\rho}}{m}} + 3\sqrt{\frac{\log \frac{4}{\delta}}{2m}}\,. \tag{4.43}$$

Proof Consider two sequences $(\rho_k)_{k\geq 1}$ and $(\epsilon_k)_{k\geq 1}$, with $\epsilon_k \in (0, 1)$. By theorem 4.4, for any fixed $k \geq 1$,

$$\Pr\left[R(h) - \widehat{R}_{\rho_k}(h) > \frac{2}{\rho_k}\mathfrak{R}_m(H) + \epsilon_k\right] \leq \exp(-2m\epsilon_k^2). \tag{4.44}$$

Choose $\epsilon_k = \epsilon + \sqrt{\frac{\log k}{m}}$, then, by the union bound,

$$\begin{aligned}
\Pr\left[\exists k\colon R(h) - \widehat{R}_{\rho_k}(h) > \frac{2}{\rho_k}\mathfrak{R}_m(H) + \epsilon_k\right] &\leq \sum_{k\geq 1} \exp(-2m\epsilon_k^2) \\
&= \sum_{k\geq 1} \exp\big[-2m(\epsilon + \sqrt{(\log k)/m})^2\big] \\
&\leq \sum_{k\geq 1} \exp(-2m\epsilon^2)\exp(-2\log k) \\
&= \big(\sum_{k\geq 1} 1/k^2\big)\exp(-2m\epsilon^2) \\
&= \frac{\pi^2}{6}\exp(-2m\epsilon^2) \leq 2\exp(-2m\epsilon^2).
\end{aligned}$$

We can choose $\rho_k = 1/2^k$. For any $\rho \in (0, 1)$, there exists $k \geq 1$ such that $\rho \in (\rho_k, \rho_{k-1}]$, with $\rho_0 = 1$. For that k, $\rho \leq \rho_{k-1} = 2\rho_k$, thus $1/\rho_k \leq 2/\rho$ and $\log k = \sqrt{\log\log_2(1/\rho_k)} \leq \sqrt{\log\log_2(2/\rho)}$. Furthermore, for any $h \in H$, $\widehat{R}_{\rho_k}(h) \leq \widehat{R}_\rho(h)$. Thus,

$$\Pr\left[\exists k\colon R(h) - \widehat{R}_\rho(h) > \frac{4}{\rho}\mathfrak{R}_m(H) + \sqrt{\frac{\log\log_2(2/\rho)}{m}} + \epsilon\right] \leq 2\exp(-2m\epsilon^2),$$

which proves the first statement. The second statement can be proven in a similar

way. ■

Combining theorem 4.3 and theorem 4.4 gives directly the following general margin bound for linear hypotheses with bounded weight vectors, presented in corollary 4.1.

Corollary 4.1
Let $H = \{\mathbf{x} \mapsto \mathbf{w} \cdot \mathbf{x} \colon \|\mathbf{w}\| \leq \Lambda\}$ *and assume that* $X \subseteq \{\mathbf{x} \colon \|\mathbf{x}\| \leq r\}$. *Fix* $\rho > 0$, *then, for any* $\delta > 0$, *with probability at least* $1 - \delta$, *for any* $h \in H$,

$$R(h) \leq \widehat{R}_\rho(h) + 2\sqrt{\frac{r^2\Lambda^2/\rho^2}{m}} + \sqrt{\frac{\log \frac{1}{\delta}}{2m}}\,. \tag{4.45}$$

As with theorem 4.4, the bound of this corollary can be generalized to hold uniformly for all $\rho > 0$ at the cost of an additional term $\sqrt{\frac{\log \log_2 \frac{2}{\rho}}{m}}$ by combining theorems 4.3 and 4.5. This generalization bound for linear hypotheses is remarkable, since it does not depend directly on the dimension of the feature space, but only on the margin. It suggests that a small generalization error can be achieved when ρ/r is large (small second term) while the empirical margin loss is relatively small (first term). The latter occurs when few points are either classified incorrectly or correctly, but with margin less than ρ.

The fact that the guarantee does not explicitly depend on the dimension of the feature space may seem surprising and appear to contradict the VC-dimension lower bounds of theorems 3.6 and 3.7. Those lower bounds show that for any learning algorithm $\mathcal{A}$ there exists a *bad* distribution for which the error of the hypothesis returned by the algorithm is $\Omega(\sqrt{d/m})$ with a non-zero probability. The bound of the corollary does not rule out such *bad* cases, however: for such bad distributions, the empirical margin loss would be large even for a relatively small margin ρ, and thus the bound of the corollary would be loose in that case.

Thus, in some sense, the learning guarantee of the corollary hinges upon the hope of a good margin value ρ: if there exists a relatively large margin value $\rho > 0$ for which the empirical margin loss is small, then a small generalization error is guaranteed by the corollary. This favorable margin situation depends on the distribution: while the learning bound is distribution-independent, the existence of a good margin is in fact distribution-dependent. A favorable margin seems to appear relatively often in applications.

The bound of the corollary gives a strong justification for margin-maximization algorithms such as SVMs. First, note that for $\rho = 1$, the margin loss can be upper bounded by the hinge loss:

$$\forall x \in \mathbb{R}, \Phi_1(x) \leq \max(1 - x, 0). \tag{4.46}$$

Using this fact, the bound of the corollary implies that with probability at least $1-\delta$, for all $h \in H = \{\mathbf{x} \mapsto \mathbf{w} \cdot \mathbf{x} \colon \|\mathbf{w}\| \leq \Lambda\}$,

$$R(h) \leq \frac{1}{m}\sum_{i=1}^{m} \xi_i + 2\sqrt{\frac{r^2\Lambda^2}{m}} + \sqrt{\frac{\log \frac{1}{\delta}}{2m}}\,, \tag{4.47}$$

where $\xi_i = \max(1 - y_i(\mathbf{w} \cdot \mathbf{x}_i), 0)$. The objective function minimized by the SVM algorithm has precisely the form of this upper bound: the first term corresponds to the slack penalty over the training set and the second to the minimization of the $\|\mathbf{w}\|$ which is equivalent to that of $\|\mathbf{w}\|^2$. Note that an alternative objective function would be based on the empirical margin loss instead of the hinge loss. However, the advantage of the hinge loss is that it is convex, while the margin loss is not.

As already pointed out, the bounds just discussed do not directly depend on the dimension of the feature space and guarantee good generalization with a favorable margin. Thus, they suggest seeking large-margin separating hyperplanes in a very high-dimensional space. In view of the form of the dual optimization problems for SVMs, determining the solution of the optimization and using it for prediction both require computing many inner products in that space. For very high-dimensional spaces, the computation of these inner products could become very costly. The next chapter provides a solution to this problem which further generalizes SVMs to non-linear separation.

4.5 Chapter notes

The maximum-margin or *optimal hyperplane* solution described in section 4.2 was introduced by Vapnik and Chervonenkis [1964]. The algorithm had limited applications, since in most tasks in practice the data is not linearly separable. In contrast, the SVM algorithm of section 4.3 for the general non-separable case, introduced by Cortes and Vapnik [1995] under the name *support-vector networks*, has been widely adopted and been shown to be effective in practice. The algorithm and its theory have had a profound impact on theoretical and applied machine learning and inspired research on a variety of topics. Several specialized algorithms have been suggested for solving the specific QP that arises when solving the SVM problem, for example the SMO algorithm of Platt [1999] (see exercise 4.4) and a variety of other decomposition methods such as those used in the LibLinear software library [Hsieh et al., 2008], and [Allauzen et al., 2010] for solving the problem when using rational kernels (see chapter 5).

Much of the theory supporting the SVM algorithm ([Cortes and Vapnik, 1995, Vapnik, 1998]), in particular the margin theory presented in section 4.4, has been adopted in the learning theory and statistics communities and applied to a variety

of other problems. The margin bound on the VC-dimension of canonical hyperplanes (theorem 4.2) is by Vapnik [1998], the proof is very similar to Novikoff's margin bound on the number of updates made by the Perceptron algorithm in the separable case. Our presentation of margin guarantees based on the Rademacher complexity follows the elegant analysis of Koltchinskii and Panchenko [2002] (see also Bartlett and Mendelson [2002], Shawe-Taylor et al. [1998]). Our proof of Talagrand's lemma 4.2 is a simpler and more concise version of a more general result given by Ledoux and Talagrand [1991, pp. 112–114]. See Höffgen et al. [1995] for hardness results related to the problem of finding a hyperplane with the minimal number of errors on a training sample.

4.6 Exercises

4.1 Soft margin hyperplanes. The function of the slack variables used in the optimization problem for soft margin hyperplanes has the form: $\xi \mapsto \sum_{i=1}^{m} \xi_i$. Instead, we could use $\xi \mapsto \sum_{i=1}^{m} \xi_i^p$, with $p > 1$.

(a) Give the dual formulation of the problem in this general case.

(b) How does this more general formulation ($p > 1$) compare to the standard setting ($p = 1$)? In the case $p = 2$ is the optimization still convex?

Sparse SVM. One can give two types of arguments in favor of the SVM algorithm: one based on the sparsity of the support vectors, another based on the notion of margin. Suppose that instead of maximizing the margin, we choose instead to maximize sparsity by minimizing the L_p norm of the vector $\boldsymbol{\alpha}$ that defines the weight vector $\mathbf{w}$, for some $p \geq 1$. First, consider the case $p = 2$. This gives the following optimization problem:

$$\begin{aligned} \min_{\boldsymbol{\alpha},b} \ & \frac{1}{2}\sum_{i=1}^{m} \alpha_i^2 + C\sum_{i=1}^{m} \xi_i \\ \text{subject to } & y_i\Big(\sum_{j=1}^{m} \alpha_j y_j \mathbf{x}_i \cdot \mathbf{x}_j + b\Big) \geq 1 - \xi_i, i \in [1,m] \\ & \xi_i, \alpha_i \geq 0, i \in [1,m]. \end{aligned} \tag{4.48}$$

(a) Show that modulo the non-negativity constraint on $\boldsymbol{\alpha}$, the problem coincides with an instance of the primal optimization problem of SVM.

(b) Derive the dual optimization of problem of (4.48).

(c) Setting $p = 1$ will induce a more sparse $\boldsymbol{\alpha}$. Derive the dual optimization in

this case.

4.2 Tighter Rademacher Bound. Derive the following tighter version of the bound of theorem 4.5: for any $\delta > 0$, with probability at least $1 - \delta$, for all $h \in H$ and $\rho \in (0, 1)$ the following holds:

$$R(h) \leq \widehat{R}_\rho(h) + \frac{2\gamma}{\rho}\mathfrak{R}_m(H) + \sqrt{\frac{\log\log_\gamma \frac{\gamma}{\rho}}{m}} + \sqrt{\frac{\log\frac{2}{\delta}}{2m}} \tag{4.49}$$

for any $\gamma > 1$.

4.3 Importance weighted SVM. Suppose you wish to use SVMs to solve a learning problem where some training data points are more important than others. More formally, assume that each training point consists of a triplet (x_i, y_i, p_i), where $0 \leq p_i \leq 1$ is the importance of the ith point. Rewrite the primal SVM constrained optimization problem so that the penalty for mis-labeling a point x_i is scaled by the priority p_i. Then carry this modification through the derivation of the dual solution.

4.4 Sequential minimal optimization (SMO). The SMO algorithm is an optimization algorithm introduced to speed up the training of SVMs. SMO reduces a (potentially) large quadratic programming (QP) optimization problem into a series of small optimizations involving only two Lagrange multipliers. SMO reduces memory requirements, bypasses the need for numerical QP optimization and is easy to implement. In this question, we will derive the update rule for the SMO algorithm in the context of the dual formulation of the SVM problem.

(a) Assume that we want to optimize equation 4.32 only over α_1 and α_2. Show that the optimization problem reduces to

$$\max_{\alpha_1,\alpha_2} \underbrace{\alpha_1 + \alpha_2 - \frac{1}{2}K_{11}\alpha_1^2 - \frac{1}{2}K_{22}\alpha_2^2 - sK_{12}\alpha_1\alpha_2 - y_1\alpha_1 v_1 - y_2\alpha_2 v_2}_{\Psi_1(\alpha_1,\alpha_2)}$$
$$\text{subject to: } 0 \leq \alpha_1, \alpha_2 \leq C \wedge \alpha_1 + s\alpha_2 = \gamma\,,$$

where $\gamma = y_1 \sum_{i=3}^m y_i\alpha_i$, $s = y_1 y_2 \in \{-1, +1\}$, $K_{ij} = (\mathbf{x}_i \cdot \mathbf{x}_j)$ and $v_i = \sum_{j=3}^m \alpha_j y_j K_{ij}$ for $i = 1, 2$.

(b) Substitute the linear constraint $\alpha_1 = \gamma - s\alpha_2$ into Ψ_1 to obtain a new objective function Ψ_2 that depends only on α_2. Show that the α_2 that minimizes Ψ_2 (without the constraints $0 \leq \alpha_1, \alpha_2 \leq C$) can be expressed as

$$\alpha_2 = \frac{s(K_{11} - K_{12})\gamma + y_2(v_1 - v_2) - s + 1}{\eta}\,,$$

where $\eta = K_{11} + K_{22} - 2K_{12}$.

(c) Show that

$$v_1 - v_2 = f(\mathbf{x}_1) - f(\mathbf{x}_2) + \alpha_2^* y_2 \eta - s y_2 \gamma (K_{11} - K_{12})$$

where $f(\mathbf{x}) = \sum_{i=1}^{m} \alpha_i^* y_i (\mathbf{x}_i \cdot \mathbf{x}) + b^*$ and α_i^* are values for the Lagrange multipliers prior to optimization over α_1 and α_2 (similarly, b^* is the previous value for the offset).

(d) Show that

$$\alpha_2 = \alpha_2^* + y_2 \frac{(y_2 - f(\mathbf{x}_2)) - (y_1 - f(\mathbf{x}_1))}{\eta} .$$

(e) For $s = +1$, define $L = \max\{0, \gamma - C\}$ and $H = \min\{C, \gamma\}$ as the lower and upper bounds on α_2. Similarly, for $s = -1$, define $L = \max\{0, -\gamma\}$ and $H = \min\{C, C - \gamma\}$. The update rule for SMO involves "clipping" the value of α_2, i.e.,

$$\alpha_2^{clip} = \begin{cases} \alpha_2 & \text{if } L < \alpha_2 < H \\ L & \text{if } \alpha_2 \leq L \\ H & \text{if } \alpha_2 \geq H \end{cases} .$$

We subsequently solve for α_1 such that we satisfy the equality constraint, resulting in $\alpha_1 = \alpha_1^* + s(\alpha_2^* - \alpha_2^{clip})$. Why is "clipping" is required? How are L and H derived for the case $s = +1$?

4.5 SVMs hands-on.

(a) Download and install the `libsvm` software library from:

`http://www.csie.ntu.edu.tw/~cjlin/libsvm/`.

(b) Download the `satimage` data set found at:

`http://www.csie.ntu.edu.tw/~cjlin/libsvmtools/datasets/`.

Merge the training and validation sets into one. We will refer to the resulting set as the training set from now on. Normalize both the training and test vectors.

(c) Consider the binary classification that consists of distinguishing class 6 from the rest of the data points. Use SVMs combined with polynomial kernels (see chapter 5) to solve this classification problem. To do so, randomly split the training data into ten equal-sized disjoint sets. For each value of the polynomial degree, $d = 1, 2, 3, 4$, plot the average cross-validation error plus or minus one standard deviation as a function of C (let the other parameters of polynomial kernels in `libsvm`, γ and c, be equal to their default values 1). Report the best

value of the trade-off constant C measured on the validation set.

(d) Let (C^*, d^*) be the best pair found previously. Fix C to be C^*. Plot the ten-fold cross-validation training and test errors for the hypotheses obtained as a function of d. Plot the average number of support vectors obtained as a function of d.

(e) How many of the support vectors lie on the margin hyperplanes?

(f) In the standard two-group classification, errors on positive or negative points are treated in the same manner. Suppose, however, that we wish to penalize an error on a negative point (false positive error) $k > 0$ times more than an error on a positive point. Give the dual optimization problem corresponding to SVMs modified in this way.

(g) Assume that k is an integer. Show how you can use `libsvm` without writing any additional code to find the solution of the modified SVMs just described.

(h) Apply the modified SVMs to the classification task previously examined and compare with your previous SVMs results for $k = 2, 4, 8, 16$.