5

Optimisation Theory

All of the inductive strategies presented in Chapter 4 have a similar form. The hypothesis function should be chosen to minimise (or maximise) a certain functional. In the case of linear learning machines (LLMs), this amounts to finding a vector of parameters that minimises (or maximises) a certain cost function, typically subject to some constraints. Optimisation theory is the branch of mathematics concerned with characterising the solutions of classes of such problems, and developing effective algorithms for finding them. The machine learning problem has therefore been converted into a form that can be analysed within the framework of optimisation theory.

Depending on the specific cost function and on the nature of the constraints, we can distinguish a number of classes of optimisation problems that are well understood and for which efficient solution strategies exist. In this chapter we will describe some of the results that apply to cases in which the cost function is a convex quadratic function, while the constraints are linear. This class of optimisation problems are called convex quadratic programmes, and it is this class that proves adequate for the task of training SVMs.

Optimisation theory will not only provide us with algorithmic techniques, but also define the necessary and sufficient conditions for a given function to be a solution. An example of this is provided by the theory of duality, which will provide us with a natural interpretation of the dual representation of LLMs presented in the previous chapters. Furthermore, a deeper understanding of the mathematical structure of solutions will inspire many specific algorithmic heuristics and implementation techniques described in Chapter 7.

5.1 Problem Formulation

The general form of the problem considered in this chapter is that of finding the maximum or minimum of a function, typically subject to some constraints. The general optimisation problem can be stated as follows:

Definition 5.1 *(Primal optimisation problem)* Given functions f, g_i, $i = 1,\dots,k$, and h_i, $i = 1,\dots,m$, defined on a domain $\Omega \subseteq \mathbb{R}^n$,

$$\begin{array}{lll} \text{minimise} & f(\mathbf{w}), & \mathbf{w} \in \Omega, \\ \text{subject to} & g_i(\mathbf{w}) \leq 0, & i = 1,\dots,k, \\ & h_i(\mathbf{w}) = 0, & i = 1,\dots,m, \end{array}$$

where $f(\mathbf{w})$ is called the *objective function*, and the remaining relations are called, respectively, the *inequality* and *equality constraints*. The optimal value of the objective function is called the *value of the optimisation problem.*

To simplify the notation we will write $\mathbf{g}(\mathbf{w}) \leq \mathbf{0}$ to indicate $g_i(\mathbf{w}) \leq 0$, $i = 1,\dots,k$. The expression $\mathbf{h}(\mathbf{w}) = \mathbf{0}$ has a similar meaning for the equality constraints.

Since maximisation problems can be converted to minimisation ones by reversing the sign of $f(\mathbf{w})$, the choice of minimisation does not represent a restriction. Similarly any constraints can be rewritten in the above form.

The region of the domain where the objective function is defined and where all the constraints are satisfied is called the *feasible region*, and we will denote it by

$$R = \{\mathbf{w} \in \Omega : \mathbf{g}(\mathbf{w}) \leq \mathbf{0}, \mathbf{h}(\mathbf{w}) = \mathbf{0}\}.$$

A solution of the optimisation problem is a point $\mathbf{w}^* \in R$ such that there exists no other point $\mathbf{w} \in R$ for which $f(\mathbf{w}) < f(\mathbf{w}^*)$. Such a point is also known as a global minimum. A point $\mathbf{w}^* \in \Omega$ is called a *local minimum* of $f(\mathbf{w})$ if $\exists \varepsilon > 0$ such that $f(\mathbf{w}) \geq f(\mathbf{w}^*)$, $\forall \mathbf{w} \in \Omega$ such that $||\mathbf{w} - \mathbf{w}^*|| < \varepsilon$.

Different assumptions on the nature of the objective function and the constraints create different optimisation problems.

Definition 5.2 An optimisation problem in which the objective function, inequality and equality constraints are all linear functions is called a *linear programme.* If the objective function is quadratic while the constraints are all linear, the optimisation problem is called a *quadratic programme.*

An inequality constraint $g_i(\mathbf{w}) \leq 0$ is said to be *active* (or tight) if the solution $\mathbf{w}^*$ satisfies $g_i(\mathbf{w}^*) = 0$, otherwise it is said to be *inactive*. In this sense, equality constraints are always active. Sometimes, quantities called *slack variables* and denoted by ξ are introduced, to transform an inequality constraint into an equality one, as follows:

$$g_i(\mathbf{w}) \leq 0 \iff g_i(\mathbf{w}) + \xi_i = 0, \text{ with } \xi_i \geq 0.$$

Slack variables associated with active constraints are equal to zero, while those for inactive constraints indicate the amount of 'looseness' in the constraint.

We will consider restricted classes of optimisation problems. First we define what are meant by a convex function and a convex set.

https://doi.org/10.1017/CBO9780511801389.007 Published online by Cambridge University Press

Definition 5.3 A real-valued function $f(\mathbf{w})$ is called *convex* for $\mathbf{w} \in \mathbb{R}^n$ if, $\forall \mathbf{w}, \mathbf{u} \in \mathbb{R}^n$, and for any $\theta \in (0,1)$,

$$f(\theta\mathbf{w} + (1-\theta)\mathbf{u}) \leq \theta f(\mathbf{w}) + (1-\theta)f(\mathbf{u}).$$

If a strict inequality holds, the function is said to be *strictly convex*. A function that is twice differentiable will be convex provided its Hessian matrix is positive semi-definite. An *affine* function is one that can be expressed in the form

$$f(\mathbf{w}) = \mathbf{A}\mathbf{w} + \mathbf{b},$$

for some matrix $\mathbf{A}$ and vector $\mathbf{b}$. Note that affine functions are convex as they have zero Hessian. A set $\Omega \subseteq \mathbb{R}^n$ is called *convex* if, $\forall \mathbf{w}, \mathbf{u} \in \Omega$, and for any $\theta \in (0,1)$, the point $(\theta\mathbf{w} + (1-\theta)\mathbf{u}) \in \Omega$.

If a function f is convex, any local minimum $\mathbf{w}^*$ of the unconstrained optimisation problem with objective function f is also a global minimum, since for any $\mathbf{u} \neq \mathbf{w}^*$, by the definition of a local minimum there exists θ sufficiently close to 1 that

$$\begin{aligned} f(\mathbf{w}^*) &\leq f(\theta\mathbf{w}^* + (1-\theta)\mathbf{u}) \\ &\leq \theta f(\mathbf{w}^*) + (1-\theta)f(\mathbf{u}). \end{aligned}$$

It follows that $f(\mathbf{w}^*) < f(\mathbf{u})$. It is this property of convex functions that renders optimisation problems tractable when the functions and sets involved are convex.

Definition 5.4 An optimisation problem in which the set Ω, the objective function and all of the constraints are convex is said to be *convex*.

For the purposes of training SVMs we can restrict ourselves to the case where the constraints are linear, the objective function is convex and quadratic and $\Omega = \mathbb{R}^n$, hence we consider convex quadratic programmes.

Optimisation theory is concerned both with describing basic properties that characterise the optimal points, and with the design of algorithms for obtaining solutions. In this chapter we will focus on the theoretical aspects, leaving algorithmic considerations to be addressed in Chapter 7. The next section will present the technique of Lagrange multipliers and its extensions, always restricted to the case of convex quadratic programmes.

5.2 Lagrangian Theory

The purpose of Lagrangian theory is to characterise the solution of an optimisation problem initially when there are no inequality constraints. The main concepts of this theory are the Lagrange multipliers and the Lagrangian function. This method was developed by Lagrange in 1797 for mechanical problems, generalising a result of Fermat from 1629. In 1951 Kuhn and Tucker further

extended the method to allow inequality constraints in what is known as Kuhn–Tucker theory. These three increasingly general results will provide all that we need to develop efficient solutions for the task of optimising SVMs. For ease of understanding we first introduce the simplest case and then go on to consider the more complex type of problems. When there are no constraints the stationarity of the objective function is sufficient to characterise the solution.

Theorem 5.5 *(Fermat) A necessary condition for* $\mathbf{w}^*$ *to be a minimum of* $f(\mathbf{w})$, $f \in C^1$, *is* $\frac{\partial f(\mathbf{w}^*)}{\partial \mathbf{w}} = \mathbf{0}$. *This condition, together with convexity of* f, *is also a sufficient condition.*

We will give one simple example of this type of optimisation taken from Chapter 3 when we considered finding the best approximation in a reproducing kernel Hilbert space.

Example 5.6 Suppose we wish to perform regression from a training set

$$S = ((\mathbf{x}_1, y_1), \ldots, (\mathbf{x}_\ell, y_\ell)) \subset (X \times Y)^\ell \subset (\mathbb{R}^n \times \mathbb{R})^\ell,$$

generated from the target function $t(\mathbf{x})$. If we assume a dual representation of the form

$$f(\mathbf{x}) = \sum_{i=1}^{\ell} \alpha_i K(\mathbf{x}_i, \mathbf{x}),$$

in Example 3.11 we showed that to minimise the RKHS norm of the error we must minimise

$$-2\sum_{i=1}^{\ell} \alpha_i y_i + \sum_{i=1}^{\ell}\sum_{j=1}^{\ell} \alpha_i \alpha_j K(\mathbf{x}_i, \mathbf{x}_j).$$

The positive semi-definiteness of the kernel K ensures that the objective function is convex. Using Theorem 5.5 we compute the derivatives with respect to α_i and set equal to zero, obtaining

$$-2y_i + 2\sum_{j=1}^{\ell} \alpha_j K(\mathbf{x}_i, \mathbf{x}_j) = 0, \; i = 1, \ldots, \ell,$$

or

$$\mathbf{G}\boldsymbol{\alpha} = \mathbf{y},$$

where we have denoted by $\mathbf{G}$ the Gram matrix with entries $\mathbf{G}_{ij} = K(\mathbf{x}_i, \mathbf{x}_j)$. Hence, the parameters $\boldsymbol{\alpha}^*$ for the solution can be obtained as

$$\boldsymbol{\alpha} = \mathbf{G}^{-1}\mathbf{y}.$$

https://doi.org/10.1017/CBO9780511801389.007 Published online by Cambridge University Press

In constrained problems, one needs to define a function, known as the Lagrangian, that incorporates information about both the objective function and the constraints, and whose stationarity can be used to detect solutions. Precisely, the Lagrangian is defined as the objective function plus a linear combination of the constraints, where the coefficients of the combination are called the Lagrange multipliers.

Definition 5.7 Given an optimisation problem with objective function $f(\mathbf{w})$, and equality constraints $h_i(\mathbf{w}) = 0$, $i = 1, \ldots, m$, we define the *Lagrangian function* as

$$L(\mathbf{w}, \boldsymbol{\beta}) = f(\mathbf{w}) + \sum_{i=1}^{m} \beta_i h_i(\mathbf{w})$$

where the coefficients β_i are called the *Lagrange multipliers.*

If a point $\mathbf{w}^*$ is a local minimum for an optimisation problem with only equality constraints, it is possible that $\frac{\partial f(\mathbf{w}^*)}{\partial \mathbf{w}} \neq \mathbf{0}$, but that the directions in which we could move to reduce f cause us to violate one or more of the constraints. In order to respect the equality constraint h_i, we must move perpendicular to $\frac{\partial h_i(\mathbf{w}^*)}{\partial \mathbf{w}}$, and so to respect all of the constraints we must move perpendicular to the subspace V spanned by

$$\left\{ \frac{\partial h_i(\mathbf{w}^*)}{\partial \mathbf{w}} : i = 1, \ldots, m \right\}.$$

If the $\frac{\partial h_i(\mathbf{w}^*)}{\partial \mathbf{w}}$ are linearly independent no legal move can change the value of the objective function, whenever $\frac{\partial f(\mathbf{w}^*)}{\partial \mathbf{w}}$ lies in the subspace V or in other words when there exist β_i such that

$$\frac{\partial f(\mathbf{w}^*)}{\partial \mathbf{w}} + \sum_{i=1}^{m} \beta_i h_i(\mathbf{w}^*) = 0.$$

This observation forms the basis of the second optimisation result concerning optimisation problems with equality constraints.

Theorem 5.8 *(Lagrange) A necessary condition for a normal point* $\mathbf{w}^*$ *to be a minimum of* $f(\mathbf{w})$ *subject to* $h_i(\mathbf{w}) = 0$, $i = 1, \ldots, m$, *with* $f, h_i \in C^1$, *is*

$$\begin{aligned} \frac{\partial L(\mathbf{w}^*, \boldsymbol{\beta}^*)}{\partial \mathbf{w}} &= \mathbf{0}, \\ \frac{\partial L(\mathbf{w}^*, \boldsymbol{\beta}^*)}{\partial \beta} &= 0, \end{aligned}$$

for some values $\boldsymbol{\beta}^*$. *The above conditions are also sufficient provided that* $L(\mathbf{w}, \boldsymbol{\beta}^*)$ *is a convex function of* $\mathbf{w}$.

The first of the two conditions gives a new system of equations, whereas the second returns the equality constraints. By imposing the conditions (jointly solving the two systems) one obtains the solution.

Example 5.9 (Largest volume box with a given surface) Let us consider the problem of finding the dimensions of the sides of a box, w, u, v whose volume is maximal and whose surface is equal to a given value c. The problem can be rewritten as

$$\begin{array}{ll} \text{minimise} & -wuv \\ \text{subject to} & wu + uv + vw = c/2. \end{array}$$

The Lagrangian of this problem is $L = -wuv + \beta(wu + uv + vw - c/2)$ and the necessary conditions for optimality are given by the constraints and by the stationarity conditions

$$\begin{aligned} \frac{\partial L}{\partial w} &= -uv + \beta(u + v) = 0, \\ \frac{\partial L}{\partial u} &= -vw + \beta(v + w) = 0, \\ \frac{\partial L}{\partial v} &= -wu + \beta(w + u) = 0. \end{aligned}$$

These conditions imply that $\beta v(w - u) = 0$ and $\beta w(u - v) = 0$, whose only non-trivial solution is $w = u = v = \sqrt{\frac{c}{6}}$. Hence since the conditions are necessary for a minimum and the trivial solutions have zero volume, the maximum volume box is a cube.

Example 5.10 (Maximum entropy distribution) The entropy of a probability distribution $\mathbf{p} = (p_1, \ldots, p_n)$ over a finite set $\{1, 2, \ldots, n\}$ is defined as $H(\mathbf{p}) = -\sum_{i=1}^{n} p_i \log p_i$, where naturally $\sum_{i=1}^{n} p_i = 1$. The distribution with maximum entropy can be found by solving an optimisation problem with the Lagrangian

$$L(\mathbf{p}, \beta) = \sum_{i=1}^{n} p_i \log p_i + \beta \left(\sum_{i=1}^{n} p_i - 1 \right)$$

over the domain $\Omega = \{\mathbf{p} : p_i \geq 0, i = 1, \ldots, n\}$. The stationarity conditions imply that $\log e p_i + \beta = 0$ for all i, indicating that all p_i need to be equal to $\frac{2^{-\beta}}{e}$. This together with the constraint gives $\mathbf{p} = (\frac{1}{n}, \ldots, \frac{1}{n})$. Since Ω is convex, the constraint is affine and the objective function is convex, having a diagonal Hessian with entries $(e p_i \ln 2)^{-1}$, this shows that the uniform distribution has the maximum entropy.

Remark 5.11 Note that if we replace the ith constraint by $h_i(\mathbf{w}) = b_i$, and consider the value of the objective function at the optimal solution $f^* = f(\mathbf{w}^*)$ as a function of b_i, then $\left[\frac{\partial f^*}{\partial b_i}\right]_{b_i=0} = \beta_i^*$. Hence the Lagrange multipliers contain information about the sensitivity of the solution to a given constraint.

Remark 5.12 Note that since the constraints are equal to zero, the value of the Lagrangian at the optimal point is equal to the value of the objective function

$$L(\mathbf{w}^*, \boldsymbol{\beta}^*) = f(\mathbf{w}^*).$$

https://doi.org/10.1017/CBO9780511801389.007 Published online by Cambridge University Press

We now consider the most general case where the optimisation problem contains both equality and inequality constraints. We first give the definition of the generalised Lagrangian.

Definition 5.13 Given an optimisation problem with domain $\Omega \subseteq \mathbb{R}^n$,

$$\begin{array}{lll} \text{minimise} & f(\mathbf{w}), & \mathbf{w} \in \Omega, \\ \text{subject to} & g_i(\mathbf{w}) \leq 0, & i = 1, \ldots, k, \\ & h_i(\mathbf{w}) = 0, & i = 1, \ldots, m, \end{array}$$

we define the *generalised Lagrangian function* as

$$\begin{aligned} L(\mathbf{w}, \boldsymbol{\alpha}, \boldsymbol{\beta}) &= f(\mathbf{w}) + \sum_{i=1}^{k} \alpha_i g_i(\mathbf{w}) + \sum_{i=1}^{m} \beta_i h_i(\mathbf{w}) \\ &= f(\mathbf{w}) + \boldsymbol{\alpha}' \mathbf{g}(\mathbf{w}) + \boldsymbol{\beta}' \mathbf{h}(\mathbf{w}). \end{aligned}$$

We can now define the Lagrangian dual problem.

Definition 5.14 The *Lagrangian dual problem* of the primal problem of Definition 5.1 is the following problem:

$$\begin{array}{ll} \text{maximise} & \theta(\boldsymbol{\alpha}, \boldsymbol{\beta}), \\ \text{subject to} & \boldsymbol{\alpha} \geq \mathbf{0}, \end{array}$$

where $\theta(\boldsymbol{\alpha}, \boldsymbol{\beta}) = \inf_{\mathbf{w} \in \Omega} L(\mathbf{w}, \boldsymbol{\alpha}, \boldsymbol{\beta})$. The value of the objective function at the optimal solution is called the *value of the problem.*

We begin by proving a theorem known as the weak duality theorem, which gives one of the fundamental relationships between the primal and dual problems and has two useful corollaries.

Theorem 5.15 *Let $\mathbf{w} \in \Omega$ be a feasible solution of the primal problem of Definition 5.1 and $(\boldsymbol{\alpha}, \boldsymbol{\beta})$ a feasible solution of the dual problem of Definition 5.14. Then $f(\mathbf{w}) \geq \theta(\boldsymbol{\alpha}, \boldsymbol{\beta})$.*

Proof From the definition of $\theta(\boldsymbol{\alpha}, \boldsymbol{\beta})$ for $\mathbf{w} \in \Omega$ we have

$$\begin{aligned} \theta(\boldsymbol{\alpha}, \boldsymbol{\beta}) &= \inf_{\mathbf{u} \in \Omega} L(\mathbf{u}, \boldsymbol{\alpha}, \boldsymbol{\beta}) \\ &\leq L(\mathbf{w}, \boldsymbol{\alpha}, \boldsymbol{\beta}) \\ &= f(\mathbf{w}) + \boldsymbol{\alpha}' \mathbf{g}(\mathbf{w}) + \boldsymbol{\beta}' \mathbf{h}(\mathbf{w}) \leq f(\mathbf{w}), \end{aligned} \tag{5.1}$$

since the feasibility of $\mathbf{w}$ implies $\mathbf{g}(\mathbf{w}) \leq \mathbf{0}$ and $\mathbf{h}(\mathbf{w}) = \mathbf{0}$, while the feasibility of $(\boldsymbol{\alpha}, \boldsymbol{\beta})$ implies $\boldsymbol{\alpha} \geq \mathbf{0}$. □

Corollary 5.16 The value of the dual is upper bounded by the value of the primal,

$$\sup \{\theta(\boldsymbol{\alpha}, \boldsymbol{\beta}) : \boldsymbol{\alpha} \geq \mathbf{0}\} \leq \inf \{f(\mathbf{w}) : \mathbf{g}(\mathbf{w}) \leq \mathbf{0}, \mathbf{h}(\mathbf{w}) = \mathbf{0}\}.$$

Corollary 5.17 If $f(\mathbf{w}^*) = \theta(\boldsymbol{\alpha}^*, \boldsymbol{\beta}^*)$, where $\boldsymbol{\alpha}^* \geq \mathbf{0}$, and $\mathbf{g}(\mathbf{w}^*) \leq \mathbf{0}$, $\mathbf{h}(\mathbf{w}^*) = \mathbf{0}$, then $\mathbf{w}^*$ and $(\boldsymbol{\alpha}^*, \boldsymbol{\beta}^*)$ solve the primal and dual problems respectively. In this case $\alpha_i^* g_i(\mathbf{w}^*) = 0$, for $i = 1, \ldots, k$.

Proof Since the values are equal the sequence of inequalities in equation (5.1) must become equalities. In particular the last inequality can only be an equality if $\alpha_i^* g_i(\mathbf{w}^*) = 0$, for all i. □

Remark 5.18 Hence, if we attempt to solve the primal and dual problems in tandem, we may be able to detect that we have reached the solution by comparing the difference between the values of the primal and dual problems. If this reduces to zero, we have reached the optimum. This approach relies on the solutions of the primal and dual having the same value, something that is not in general guaranteed. The difference between the values of the primal and dual problems is known as the *duality gap.*

Another way of detecting the absence of a duality gap is the presence of a *saddle point.* A saddle point of the Lagrangian function for the primal problem is a triple

$$(\mathbf{w}^*, \boldsymbol{\alpha}^*, \boldsymbol{\beta}^*), \text{ with } \mathbf{w}^* \in \Omega, \boldsymbol{\alpha}^* \geq \mathbf{0},$$

satisfying the additional property that

$$L(\mathbf{w}^*, \boldsymbol{\alpha}, \boldsymbol{\beta}) \leq L(\mathbf{w}^*, \boldsymbol{\alpha}^*, \boldsymbol{\beta}^*) \leq L(\mathbf{w}, \boldsymbol{\alpha}^*, \boldsymbol{\beta}^*),$$

for all $\mathbf{w} \in \Omega$, $\boldsymbol{\alpha} \geq \mathbf{0}$. Note that $\mathbf{w}$ here is not required to satisfy the equality or inequality constraints.

Theorem 5.19 *The triple* $(\mathbf{w}^*, \boldsymbol{\alpha}^*, \boldsymbol{\beta}^*)$ *is a saddle point of the Lagrangian function for the primal problem, if and only if its components are optimal solutions of the primal and dual problems and there is no duality gap, the primal and dual problems having the value*

$$f(\mathbf{w}^*) = \theta(\boldsymbol{\alpha}^*, \boldsymbol{\beta}^*).$$

We will now quote the strong duality theorem, which guarantees that the dual and primal problems have the same value for the optimisation problems that we will be considering.

Theorem 5.20 *(Strong duality theorem) Given an optimisation problem with convex domain* $\Omega \subseteq \mathbb{R}^n$,

$$\begin{array}{lll} \textit{minimise} & f(\mathbf{w}), & \mathbf{w} \in \Omega, \\ \textit{subject to} & g_i(\mathbf{w}) \leq 0, & i = 1, \ldots, k, \\ & h_i(\mathbf{w}) = 0, & i = 1, \ldots, m, \end{array}$$

where the g_i *and* h_i *are affine functions, that is*

$$\mathbf{h}(\mathbf{w}) = \mathbf{A}\mathbf{w} - \mathbf{b},$$

for some matrix $\mathbf{A}$ *and vector* $\mathbf{b}$*, the duality gap is zero.*

We are now in a position to give the Kuhn–Tucker theorem giving conditions for an optimum solution to a general optimisation problem.

Theorem 5.21 *(Kuhn–Tucker) Given an optimisation problem with convex domain* $\Omega \subseteq \mathbb{R}^n$,

$$\begin{array}{lll} \textit{minimise} & f(\mathbf{w}), & \mathbf{w} \in \Omega, \\ \textit{subject to} & g_i(\mathbf{w}) \leq 0, & i = 1, \ldots, k, \\ & h_i(\mathbf{w}) = 0, & i = 1, \ldots, m, \end{array}$$

with $f \in C^1$ *convex and* g_i, h_i *affine, necessary and sufficient conditions for a normal point* $\mathbf{w}^*$ *to be an optimum are the existence of* $\boldsymbol{\alpha}^*, \boldsymbol{\beta}^*$ *such that*

$$\begin{aligned} \frac{\partial L(\mathbf{w}^*, \boldsymbol{\alpha}^*, \boldsymbol{\beta}^*)}{\partial \mathbf{w}} &= \mathbf{0}, \\ \frac{\partial L(\mathbf{w}^*, \boldsymbol{\alpha}^*, \boldsymbol{\beta}^*)}{\partial \beta} &= 0, \\ \alpha_i^* g_i(\mathbf{w}^*) &= 0, \; i = 1, \ldots, k, \\ g_i(\mathbf{w}^*) &\leq 0, \; i = 1, \ldots, k, \\ \alpha_i^* &\geq 0, \; i = 1, \ldots, k. \end{aligned}$$

Remark 5.22 The third relation is known as Karush–Kuhn–Tucker complementarity condition. It implies that for active constraints, $\alpha_i^* \geq 0$, whereas for inactive constraints $\alpha_i^* = 0$. Furthermore, it is possible to show that for active constraints again changing the constraint to be b_i in place of 0, $\alpha_i^* = \left[\frac{\partial f^*}{\partial b_i}\right]_{b_i=0}$, so that the Lagrange multiplier represents the sensitivity of the optimal value to the constraint. Perturbing inactive constraints has no effect on the solution of the optimisation problem.

Remark 5.23 One way to interpret the above results is that a solution point can be in one of two positions with respect to an inequality constraint, either in the interior of the feasible region, with the constraint inactive, or on the boundary defined by that constraint with the constraint active. In the first case, the conditions for optimality for that constraint are given by Fermat's theorem, so the α_i need to be zero. In the second case, one can use Lagrange's theorem with a non-zero α_i. So the KKT conditions say that either a constraint is active, meaning $g_i(\mathbf{w}^*) = 0$, or the corresponding multiplier satisfies $\alpha_i^* = 0$. This is summarised in the equation $g_i(\mathbf{w}^*)\alpha_i^* = 0$.

5.3 Duality

Lagrangian treatment of convex optimisation problems leads to an alternative dual description, which often turns out to be easier to solve than the primal problem since handling inequality constraints directly is difficult. The dual problem is obtained by introducing Lagrange multipliers, also called the dual

https://doi.org/10.1017/CBO9780511801389.007 Published online by Cambridge University Press

variables. Dual methods are based on the idea that the dual variables are the fundamental unknowns of the problem.

We can transform the primal into a dual by simply setting to zero the derivatives of the Lagrangian with respect to the primal variables, and substituting the relations so obtained back into the Lagrangian, hence removing the dependence on the primal variables. This corresponds to explicitly computing the function

$$\theta(\boldsymbol{\alpha}, \boldsymbol{\beta}) = \inf_{\mathbf{w} \in \Omega} L(\mathbf{w}, \boldsymbol{\alpha}, \boldsymbol{\beta}).$$

The resulting function contains only dual variables and must be maximised under simpler constraints. This strategy will be adopted in subsequent chapters and has become one of the standard techniques in the theory of Support Vector Machines. The pleasing feature of the resulting expression for the primal variables is that it matches exactly the dual representation introduced in Chapter 2 and so will seem very natural in the context in which we will be using it. Hence, the use of dual representations in Support Vector Machines not only allows us to work in high dimensional spaces as indicated in Chapter 3, but also paves the way for algorithmic techniques derived from optimisation theory. As a further example the duality gap can be used as a convergence criterion for iterative techniques. A number of further consequences will flow from the convex quadratic programmes that arise in SVM optimisation. The Karush–Kuhn–Tucker complementarity conditions imply that only the active constraints will have non-zero dual variables, meaning that for certain optimisations the actual number of variables involved may be significantly fewer than the full training set size. We will see that the term *support vector* refers to those examples for which the dual variables are non-zero.

Example 5.24 (Quadratic programme) We demonstrate the practical use of duality by applying it to the important special case of a quadratic objective function.

$$\begin{array}{ll} \text{minimise} & \frac{1}{2}\mathbf{w}'\mathbf{Q}\mathbf{w} - \mathbf{k}'\mathbf{w}, \\ \text{subject to} & \mathbf{X}\mathbf{w} \leq \mathbf{c}, \end{array}$$

where $\mathbf{Q}$ is a positive definite $n \times n$ matrix, $\mathbf{k}$ is an n-vector; $\mathbf{c}$ an m-vector, $\mathbf{w}$ the unknown, and $\mathbf{X}$ an $m \times n$ matrix. Assuming that the feasible region is not empty, this problem can be rewritten as

$$\max_{\boldsymbol{\alpha} \geq 0} \left(\min_{\mathbf{w}} \left(\frac{1}{2}\mathbf{w}'\mathbf{Q}\mathbf{w} - \mathbf{k}'\mathbf{w} + \boldsymbol{\alpha}'(\mathbf{X}\mathbf{w} - \mathbf{c}) \right) \right).$$

The minimum over $\mathbf{w}$ is unconstrained, and is attained at $\mathbf{w} = \mathbf{Q}^{-1}(\mathbf{k} - \mathbf{X}'\boldsymbol{\alpha})$. Resubstituting this back in the original problem, one obtains the dual:

$$\begin{array}{ll} \text{maximise} & -\frac{1}{2}\boldsymbol{\alpha}'\mathbf{P}\boldsymbol{\alpha} - \boldsymbol{\alpha}'\mathbf{d} - \frac{1}{2}\mathbf{k}'\mathbf{Q}\mathbf{k}, \\ \text{subject to} & \boldsymbol{\alpha} \geq \mathbf{0}, \end{array}$$

where $\mathbf{P} = \mathbf{X}\mathbf{Q}^{-1}\mathbf{X}'$, and $\mathbf{d} = \mathbf{c} - \mathbf{X}\mathbf{Q}^{-1}\mathbf{k}$. Thus, the dual of a quadratic program is another quadratic programme but with simpler constraints.

https://doi.org/10.1017/CBO9780511801389.007 Published online by Cambridge University Press

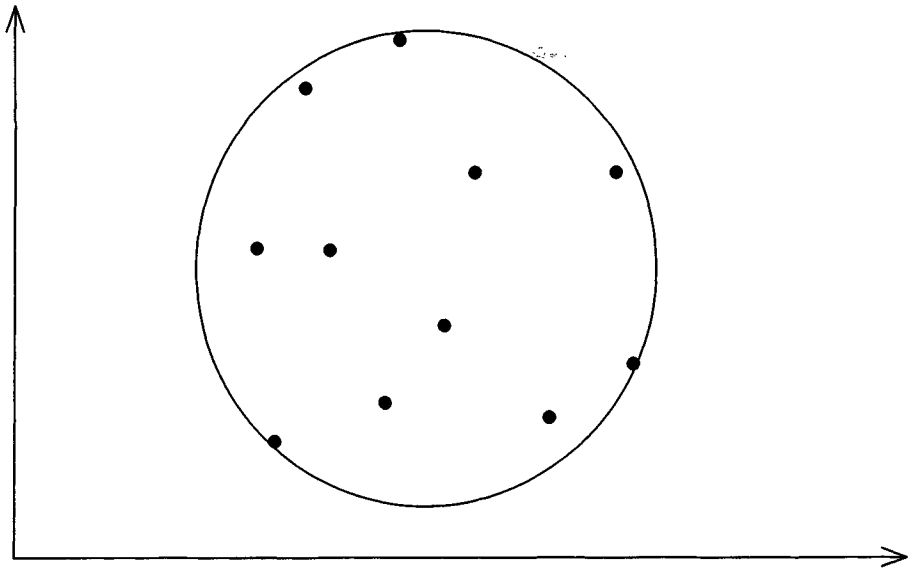

Figure 5.1: Example of a minimal enclosing sphere for a set of points in two dimensions

5.4 Exercises

1. A ball centred at a point $\mathbf{v}$ of radius R is the set

$$B_R(\mathbf{v}) = \{\mathbf{u} : \|\mathbf{u} - \mathbf{v}\| \leq R\}.$$

Express the problem of finding the ball of smallest radius that contains a given set

$$S = \{\mathbf{x}_1, \ldots, \mathbf{x}_\ell\}$$

of vectors as an optimisation problem. See Figure 5.1 for a simple two dimensional example. Convert the problem derived to the dual form, hence showing that the solution can be expressed as a linear combination of the set S and can be solved in a kernel-induced feature space.

2. The convex hull of a set

$$T = \{\mathbf{x}_1, \ldots, \mathbf{x}_\ell\}$$

is the set of all convex combinations of the points in T. Given a linearly separable training set $S = S^+ \cup S^-$ of positive and negative examples, express the problem of finding points $\mathbf{x}^+$ and $\mathbf{x}^-$ in the convex hulls of S^+ and S^- for which the distance $\|\mathbf{x}^+ - \mathbf{x}^-\|$ is minimal as an optimisation problem. Note that this distance is twice the margin of the training set S.

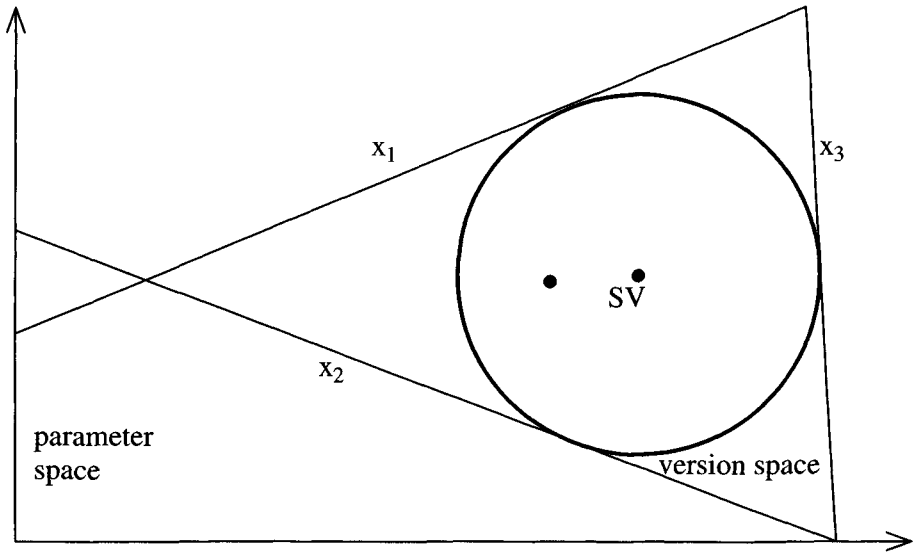

Figure 5.2: The version space for a linear learning machine

3. Consider the parameter space of weight vectors for a linear learning machine. Each point of this space corresponds to one hypothesis for fixed bias. Each training example $\mathbf{x}$ gives rise to a hyperplane in this space defined by the equation

$$\langle \mathbf{w} \cdot \mathbf{x} \rangle = 0.$$

The situation for a two dimensional weight vector is shown in Figure 5.2 for three examples. Each hyperplane divides the space into those hypotheses giving the correct classification and those giving an incorrect classification. The region of hypotheses which correctly classify the whole training set is sometimes referred to as the version space. In Figure 5.2 this is the central triangle. Express the problem of finding the centre of the largest sphere completely contained in the version space, that is the point SV in Figure 5.2. Note that the point is distinct from the centre of mass of the version space also shown in the figure. Convert the problem to the dual form.

5.5 Further Reading and Advanced Topics

The theory of optimisation dates back to the work of Fermat, who formulated the result on stationarity for unconstrained problems in the 17th century. The extension to the constrained case was made by Lagrange in 1788, for the case of equality constraints. It was not until 1951 that the theory was generalised

to the case of inequality constraints by Kuhn and Tucker [77], giving rise to the modern theory of convex optimisation. Karush had already described the optimality conditions in his dissertation in 1939 [69] and this is why the conditions arising from the Kuhn–Tucker theorem are usually referred to as the Karush–Kuhn–Tucker (KKT) conditions.

In the following years, considerable work was done by Wolfe, Mangasarian, Duran, and others, to extend the duality results known for linear programming to the convex programming case (see for example the introduction of [80]). The diffusion of computers in the 1960s greatly increased the interest in what was then known as mathematical programming, which studies the solution of problems by (usually linear or quadratic) optimisation methods.

The use of optimisation in machine learning was pioneered by Olvi Mangasarian (for example, [84], [85], [87]) with his work on linear programming machines. See also [14] by Bennett et al. for a very nice discussion of linear and quadratic optimisation techniques applied to pattern recognition. Mangasarian took his ideas to their extreme, designing algorithms that can perform data mining on datasets of hundreds of thousands of points (see Section 7.8 for more references). The perceptron algorithm described in Chapter 2 can also be regarded as a simple optimisation procedure, searching for a feasible point given a set of linear constraints specified by the data. But rather than picking any point in the feasible region (an ill-posed problem) one could choose to pick some specific point satisfying some extremal property like the ones discussed in Chapter 4 such as being maximally distant from the boundaries of the feasible region. This type of consideration will lead to the development of Support Vector Machines in the next chapter. Mangasarian's early work mainly focused on minimising the 1-norm of the solution $\mathbf{w}$. Notice finally that the application of optimisation ideas to problems like least squares regression (discussed in Chapter 2) was already a use of such concepts in machine learning.

Optimisation theory is a well-developed and quite stable field, and the standard results summarised in this chapter can be found in any good textbook. A particularly readable and comprehensive text on optimisation theory is [11]; also the classic books [80], [41] and [86] provide good introductions. Optimisation theory usually also includes the algorithmic techniques needed to solve the problem practically. We will address this issue in Chapter 7. The important contribution of optimisation theory to the theory of Support Vector Machines, however, lies not on the algorithmic side, but rather in its providing a mathematical characterisation of the solutions, via the KKT conditions, giving a mathematical meaning to the dual variables (introduced in Chapter 2) and to the margin slack vector (introduced in Chapter 4), and generally in providing a geometrical intuition for the dual problem.

Exercises 1 concerning the centre of the smallest ball containing the data is relevant if we wish to minimise the estimate of the fat shattering dimension of a given set of points by optimally shifting the origin. This problem was first studied in [128], while Exercise 2 concerning the distance between convex hulls is discussed in [14], and provides an efficient way to characterise the maximal

margin hyperplane. This is discussed in [73] where it is used to motivate an interesting algorithm (see Section 7.8 for more references).

These references are also given on the website **www.support-vector.net**, which will be kept up to date with new work, pointers to software and papers that are available on-line.