Algorithm 6.5 Predictor Corrector Method

Require: x_0, Precision ϵ
  Set $i = 0$
  repeat
    Expand f into $f(x_i) + f_{\text{simple}}(\xi, x_i) + T(\xi, x_i)$.
Predictor Solve $f(x_i) + f_{\text{simple}}(\xi^{\text{pred}}, x_i) = 0$ for ξ^{pred}.
Corrector Solve $f(x_i) + f_{\text{simple}}(\xi^{\text{corr}}, x_i) + T(\xi^{\text{pred}}, x_i) = 0$ for ξ^{corr}.
    $x_{i+1} = x_i + \xi^{\text{corr}}$.
    $i = i + 1$.
  until $|f(x_i)| \leq \epsilon$
Output: x_i

where $f_{\text{simple}}(\xi, x)$ contains the simple, possibly low order, part of f, and $T(\xi, x)$ the higher order terms, such that $f_{\text{simple}}(0, x) = T(0, x) = 0$. While in the previous example we introduced higher order terms into f that were not present before (f is only quadratic), usually such terms will already exist anyway. Hence the corrector step will just eliminate additional lower order terms without too much additional error in the approximation.
We will encounter such methods for instance in the context of interior point algorithms (Section 6.4), where we have to solve a set of quadratic equations.

6.3 Constrained Problems

After this digression on unconstrained optimization problems, let us return to constrained optimization, which makes up the main body of the problems we will have to deal with in learning (e.g., quadratic or general convex programs for Support Vector Machines). Typically, we have to deal with problems of type (6.6). For convenience we repeat the problem statement:

$$\begin{array}{ll} \underset{x}{\text{minimize}} & f(x) \\ \text{subject to} & c_i(x) \leq 0 \text{ for all } i \in [n]. \end{array} \tag{6.37}$$

Here f and c_i are convex functions and $n \in \mathbb{N}$. In some cases[3], we additionally have *equality* constraints $e_j(x) = 0$ for some $j \in [n']$. Then the optimization problem can be written as

$$\begin{array}{ll} \underset{x}{\text{minimize}} & f(x), \\ \text{subject to} & c_i(x) \leq 0 \text{ for all } i \in [n], \\ & e_j(x) = 0 \text{ for all } j \in [n']. \end{array} \tag{6.38}$$

3. Note that it is common practice in Support Vector Machines to write c_i as positivity constraints by using concave functions. This can be fixed by a sign change, however.

Before we start minimizing f, we have to discuss what optimality means in this case. Clearly $f'(x) = 0$ is too restrictive a condition. For instance, f' could point into a direction which is forbidden by the constraints c_i and e_i. Then we could have optimality, even though $f' \neq 0$. Let us analyze the situation in more detail.

6.3.1 Optimality Conditions

We start with optimality conditions for optimization problems which are independent of their differentiability. While it is fairly straightforward to state *sufficient* optimality conditions for arbitrary functions f and c_i, we will need convexity and "reasonably nice" constraints (see Lemma 6.23) to state *necessary* conditions. This is not a major concern, since for practical applications, the constraint qualification criteria are almost always satisfied, and the functions themselves are usually convex and differentiable. Much of the reasoning in this section follows [345], which should also be consulted for further references and detail.

Some of the most important sufficient criteria are the Kuhn-Tucker[4] saddle point conditions [312]. As indicated previously, they are independent of assumptions on convexity or differentiability of the constraints c_i or objective function f.

Theorem 6.21 (Kuhn-Tucker Saddle Point Condition [312, 345]) *Assume an optimization problem of the form (6.37), where $f : \mathbb{R}^m \to \mathbb{R}$ and $c_i : \mathbb{R}^m \to \mathbb{R}$ for $i \in [n]$ are arbitrary functions, and a Lagrangian*

Lagrangian

$$L(x, \alpha) := f(x) + \sum_{i=1}^{n} \alpha_i c_i(x) \text{ where } \alpha_i \geq 0. \tag{6.39}$$

If a pair of variables $(\bar{x}, \bar{\alpha})$ with $\bar{x} \in \mathbb{R}^n$ and $\bar{\alpha}_i \geq 0$ for all $i \in [n]$ exists, such that for all $x \in \mathbb{R}^m$ and $\alpha \in [0, \infty)^n$,

$$L(\bar{x}, \alpha) \leq L(\bar{x}, \bar{\alpha}) \leq L(x, \bar{\alpha}) \text{ (Saddle Point)} \tag{6.40}$$

then $\bar{x}$ is a solution to (6.37).

The parameters α_i are called Lagrange multipliers. As described in the later chapters, they will become the coefficients in the kernel expansion in SVM.

Proof The proof follows [345]. Denote by $(\bar{x}, \bar{\alpha})$ a pair of variables satisfying (6.40). From the first inequality it follows that

$$\sum_{i=1}^{n} (\alpha_i - \bar{\alpha}_i) c_i(\bar{x}) \leq 0. \tag{6.41}$$

Since we are free to choose $\alpha_i \geq 0$, we can see (by setting all but one of the terms α_i to $\bar{\alpha}_i$ and the remaining one to $\alpha_i = \bar{\alpha}_i + 1$) that $c_i(x) \leq 0$ for all $i \in [n]$. This shows that $\bar{x}$ satisfies the constraints, i.e. it is feasible.

4. An earlier version is due to Karush [283]. This is why often one uses the abbreviation KKT (Karush-Kuhn-Tucker) rather than KT to denote the optimality conditions.

Additionally, by setting one of the α_i to 0, we see that $\bar{\alpha}_i c_i(\bar{x}) \geq 0$. The only way to satisfy this is by having

$$\bar{\alpha}_i c_i(\bar{x}) = 0 \text{ for all } i \in [n]. \tag{6.42}$$

Eq. (6.42) is often referred to as the KKT condition [283, 312]. Finally, combining (6.42) and $c_i(x) \leq 0$ with the second inequality in (6.40) yields $f(\bar{x}) \leq f(x)$ for all feasible x. This proves that $\bar{x}$ is optimal. ■

We can immediately extend Theorem 6.21 to accommodate equality constraints by splitting them into the conditions $e_i(x) \leq 0$ and $e_i(x) \geq 0$. We obtain:

Theorem 6.22 (Equality Constraints) *Assume an optimization problem of the form (6.38), where $f, c_i, e_j : \mathbb{R}^m \to \mathbb{R}$ for $i \in [n]$ and $j \in [n']$ are arbitrary functions, and a Lagrangian*

$$L(x, \alpha) := f(x) + \sum_{i=1}^{n} \alpha_i c_i(x) + \sum_{j=1}^{n'} \beta_j e_j(x) \text{ where } \alpha_i \geq 0 \text{ and } \beta_j \in \mathbb{R}. \tag{6.43}$$

If a set of variables $(\bar{x}, \bar{\alpha}, \bar{\beta})$ with $\bar{x} \in \mathbb{R}^m$, $\bar{\alpha} \in [0, \infty)$, and $\bar{\beta} \in \mathbb{R}^{n'}$ exists such that for all $x \in \mathbb{R}^m$, $\alpha \in [0, \infty)^n$, and $\beta \in \mathbb{R}^{n'}$,

$$L(\bar{x}, \alpha, \beta) \leq L(\bar{x}, \bar{\alpha}, \bar{\beta}) \leq L(x, \bar{\alpha}, \bar{\beta}), \tag{6.44}$$

then $\bar{x}$ is a solution to (6.38).

Now we determine when the conditions of Theorem 6.21 are necessary. We will see that convexity and sufficiently "nice" constraints are needed for (6.40) to become a necessary condition. The following lemma (see [345]) describes three *constraint qualifications*, which will turn out to be exactly what we need.

Feasible Region

Lemma 6.23 (Constraint Qualifications) *Denote by $\mathcal{X} \subset \mathbb{R}^m$ a convex set, and by $c_1, \ldots, c_n : \mathcal{X} \to \mathbb{R}$ n convex functions defining a feasible region by*

$$X := \{x | x \in \mathcal{X} \text{ and } c_i(x) \leq 0 \text{ for all } i \in [n]\}. \tag{6.45}$$

Equivalence Between Constraint Qualifications

Then the following additional conditions on c_i are connected by $(i) \iff (ii)$ and $(iii) \implies (i)$.

(i) There exists an $x \in \mathcal{X}$ such that for all $i \in [n]$ $c_i(x) < 0$ (Slater's condition [500]).

(ii) For all nonzero $\alpha \in [0, \infty)^n$ there exists an $x \in \mathcal{X}$ such that $\sum_{i=1}^{n} \alpha_i c_i(x) \leq 0$ (Karlin's condition [281]).

(iii) The feasible region X contains at least two distinct elements, and there exists an $x \in X$ such that all c_i are strictly convex at x wrt. X (Strict constraint qualification).

The connection $(i) \iff (ii)$ is also known as the Generalized Gordan Theorem [164]. The proof can be skipped if necessary. We need an auxiliary lemma which we state without proof (see [345, 435] for details).

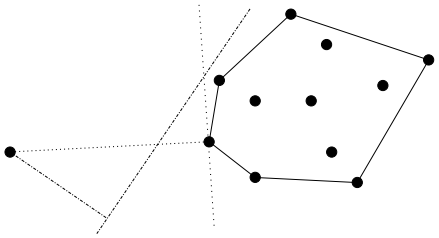

Figure 6.7 Two hyperplanes (and their normal vectors) separating the convex hull of a finite set of points from the origin.

Lemma 6.24 (Separating Hyperplane Theorem) *Denote by $X \in \mathbb{R}^m$ a convex set not containing the origin 0. Then there exists a hyperplane with normal vector $\alpha \in \mathbb{R}^m$ such that $\alpha^\top x > 0$ for all $x \in X$.*

See also Figure 6.7.

Proof of Lemma 6.23. We prove $\{(i) \iff (ii)\}$ by showing $\{(i) \implies (ii)\}$ and $\{$ not $(i) \implies$ not $(ii)\}$.

$(i) \implies (ii)$ For a point $x \in X$ with $c_i(x) < 0$, for all $i \in [n]$ we have that $\alpha_i c_i(x) \geq 0$ implies $\alpha_i = 0$.

$\overline{(i)} \implies \overline{(ii)}$ Assume that there is no x with $c_i(x) < 0$ for all $i \in [n]$. Hence the set

$$\Gamma := \{\gamma | \gamma \in \mathbb{R}^n \text{ and there exists some } x \in X \text{ with } \gamma_i > c_i(x) \text{ for all } i \in [n]\} \tag{6.46}$$

is convex and does not contain the origin. The latter follows directly from the assumption. For the former take $\gamma, \gamma' \in \Gamma$ and $\lambda \in (0,1)$ to obtain

$$\lambda\gamma_i + (1-\lambda)\gamma_i' > \lambda c_i(x) + (1-\lambda)c_i(x') \geq c_i(\lambda x + (1-\lambda)x'). \tag{6.47}$$

Now by Lemma 6.24, there exists some $\alpha \in \mathbb{R}^n$ such that $\alpha^\top \gamma \geq 0$ and $\|\alpha\|^2 = 1$ for all $\gamma \in \Gamma$. Since each of the γ_i for $\gamma \in \Gamma$ can be arbitrarily large (with respect to the other coordinates), we conclude $\alpha_i \geq 0$ for all $i \in [n]$.
Denote by $\delta := \inf_{x \in X} \sum_{i=1}^n \alpha_i c_i(x)$ and by $\delta' := \inf_{\gamma \in \Gamma} \alpha^\top \gamma$. One can see that by construction $\delta = \delta'$. By Lemma 6.24 α was chosen such that $\delta' \geq 0$, and hence $\delta \geq 0$. This contradicts (ii), however, since it implies the existence of a suitable α with $\alpha_i c_i(x) \geq 0$ for all x.

$(iii) \implies (i)$ Since X is convex we get for all c_i and for any $\lambda \in (0,1)$:

$$\lambda x + (1-\lambda)x' \in X \text{ and } 0 \geq \lambda c_i(x) + (1-\lambda)c_i(x') > c_i(\lambda x + (1-\lambda)x'). \tag{6.48}$$

This shows that $\lambda x + (1-\lambda)x'$ satisfies (i) and we are done.

■

We proved Lemma 6.23 as it provides us with a set of constraint qualifications (conditions on the constraints) that allow us to determine cases where the KKT saddle point conditions are both *necessary* and sufficient. This is important, since we will use the KKT conditions to transform optimization problems into their duals, and solve the latter numerically. For this approach to be valid, however, we must ensure that we do not change the solvability of the optimization problem.

Theorem 6.25 (Necessary KKT Conditions [312, 553, 281]) *Under the assumptions and definitions of Theorem 6.21 with the additional assumption that f and c_i are convex on the convex set $X \subseteq \mathbb{R}^m$ (containing the set of feasible solutions as a subset) and that c_i satisfy one of the constraint qualifications of Lemma 6.23, the saddle point criterion (6.40) is necessary for optimality.*

Proof Denote by $\bar{x}$ the solution to (6.37), and by X' the set

$$X' := X \cap \{x | x \in \mathcal{X} \text{ with } f(x) - f(\bar{x}) \leq 0 \text{ and } c_i(x) \leq 0 \text{ for all } i \in [n]\}. \tag{6.49}$$

By construction $\bar{x} \in X'$. Furthermore, there exists no $x' \in X'$ such that all inequality constraints including $f(x) - f(\bar{x})$ are satisfied as *strict* inequalities (otherwise $\bar{x}$ would not be optimal). In other words, X' violates Slater's conditions (i) of Lemma 6.23 (where both $\big(f(x) - f(\bar{x})\big)$ and $c(x)$ *together* play the role of $c_i(x)$), and thus also Karlin's conditions (ii). This means that there exists a nonzero vector $(\bar{\alpha}_0, \bar{\alpha}) \in \mathbb{R}^{n+1}$ with nonnegative entries such that

$$\bar{\alpha}_0(f(x) - f(\bar{x})) + \sum_{i=1}^{n} \bar{\alpha}_i c_i(x) \geq 0 \text{ for all } x \in \mathcal{X}. \tag{6.50}$$

In particular, for $x = \bar{x}$ we get $\sum_{i=1}^{n} \bar{\alpha}_i c_i(\bar{x}) \geq 0$. In addition, since $\bar{x}$ is a solution to (6.37), we have $c_i(\bar{x}) \leq 0$. Hence $\sum_{i=1}^{n} \bar{\alpha}_i c_i(\bar{x}) = 0$. This allows us to rewrite (6.50) as

$$\bar{\alpha}_0 f(x) + \sum_{i=1}^{n} \bar{\alpha}_i c_i(x) \geq \bar{\alpha}_0 f(\bar{x}) + \sum_{i=1}^{n} \bar{\alpha}_i c_i(\bar{x}). \tag{6.51}$$

This looks almost like the first inequality of (6.40), except for the $\bar{\alpha}_0$ term (which we will return to later). But let us consider the second inequality first.

Again, since $c_i(\bar{x}) \leq 0$ we have $\sum_{i=1}^{n} \alpha_i c_i(\bar{x}) \leq 0$ for all $\alpha_i \geq 0$. Adding $\bar{\alpha}_0 f(\bar{x})$ on both sides of the inequality and $\sum_{i=1}^{n} \bar{\alpha}_i c_i(\bar{x})$ on the rhs yields

$$\bar{\alpha}_0 f(\bar{x}) + \sum_{i=1}^{n} \bar{\alpha}_i c_i(\bar{x}) \geq \bar{\alpha}_0 f(\bar{x}) + \sum_{i=1}^{n} \alpha_i c_i(\bar{x}). \tag{6.52}$$

This is almost all we need for the first inequality of (6.40) .[5] If $\bar{\alpha}_0 > 0$ we can divide (6.51) and (6.52) by $\bar{\alpha}_0$ and we are done.

When $\bar{\alpha}_0 = 0$, then this implies the existence of $\bar{\alpha} \in \mathbb{R}^n$ with nonnegative entries satisfying $\sum_{i=1}^{n} \bar{\alpha}_i c_i(x) \geq 0$ for all $x \in X$. This contradicts Karlin's constraint qualification condition (ii), which allows us to rule out this case. ∎

6.3.2 Duality and KKT-Gap

Now that we have formulated necessary and sufficient optimality conditions (Theorem 6.21 and 6.25) under quite general circumstances, let us put them to practical

5. The two inequalities (6.51) and (6.52) are also known as the Fritz-John saddle point necessary optimality conditions [269], which play a similar role as the saddle point conditions for the Lagrangian (6.39) of Theorem 6.21.

use for convex differentiable optimization problems. We first derive a more practically useful form of Theorem 6.21. Our reasoning is as follows: eq. (6.40) implies that $L(\bar{x}, \bar{\alpha})$ is a *saddle point* in terms of $(\bar{x}, \bar{\alpha})$. Hence, all we have to do is write the saddle point conditions in the form of derivatives.

Theorem 6.26 (KKT for Differentiable Convex Problems [312]) *A solution to the optimization problem (6.37) with convex, differentiable f, c_i is given by $\bar{x}$, if there exists some $\bar{\alpha} \in \mathbb{R}^n$ with $\alpha_i \geq 0$ for all $i \in [n]$ such that the following conditions are satisfied:*

Primal and Dual Feasibility

$$\partial_x L(\bar{x}, \bar{\alpha}) = \partial_x f(\bar{x}) + \sum_{i=1}^{n} \bar{\alpha}_i \partial_x c_i(\bar{x}) = 0 \text{ (Saddle Point in } \bar{x}\text{)}, \tag{6.53}$$

$$\partial_{\alpha_i} L(\bar{x}, \bar{\alpha}) = c_i(\bar{x}) \leq 0 \text{ (Saddle Point in } \bar{\alpha}\text{)}, \tag{6.54}$$

$$\sum_{i=1}^{n} \bar{\alpha}_i c_i(\bar{x}) = 0 \text{ (Vanishing KKT-Gap)}. \tag{6.55}$$

Proof The easiest way to prove Theorem 6.26 is to show that for any $x \in X$, we have $f(x) - f(\bar{x}) \geq 0$. Due to convexity we may linearize and obtain

$$f(x) - f(\bar{x}) \geq (\partial_x f(\bar{x}))^\top (x - \bar{x}) \tag{6.56}$$

$$= -\sum_{i=1}^{n} \bar{\alpha}_i \left(\partial_x c_i(\bar{x})\right)^\top (x - \bar{x}) \tag{6.57}$$

$$\geq -\sum_{i=1}^{n} \bar{\alpha}_i (c_i(x) - c_i(\bar{x})) \tag{6.58}$$

$$= -\sum_{i=1}^{n} \bar{\alpha}_i c_i(x) \geq 0. \tag{6.59}$$

Here we used the convexity and differentiability of f to arrive at the rhs of (6.56) and (6.58). To obtain (6.57) we exploited the fact that at the saddle point $\partial_x f(\bar{x})$ can be replaced by the corresponding expansion in $\partial_x c_i(\bar{x})$; thus we used (6.53). Finally, for (6.59) we used the fact that the KKT gap vanishes at the optimum (6.55) and that the constraints are satisfied (6.54). ∎

Optimization by Constraint Satisfaction

In other words, we may solve a convex optimization problem by finding $(\bar{x}, \bar{\alpha})$ that satisfy the conditions of Theorem 6.26. Moreover, these conditions, together with the constraint qualifications of Lemma 6.23, ensure necessity.

Note that we transformed the problem of minimizing functions into one of solving a set of equations, for which several numerical tools are readily available. This is exactly how interior point methods work (see Section 6.4 for details on how to implement them). Necessary conditions on the constraints similar to those discussed previously can also be formulated (see [345] for a detailed discussion).

The other consequence of Theorem 6.26, or rather of the definition of the Lagrangian $L(x, \alpha)$, is that we may bound $f(\bar{x}) = L(\bar{x}, \bar{\alpha})$ from above and below *without* explicit knowledge of $f(\bar{x})$.

Theorem 6.27 (KKT-Gap) *Assume an optimization problem of type (6.37), where both f and c_i are convex and differentiable. Denote by $\bar{x}$ its solution. Then for any set of variables*

(x, α) *with* $\alpha_i \geq 0$, *and for all* $i \in [n]$ *satisfying*

$$\partial_x L(x, \alpha) = 0, \tag{6.60}$$

$$\partial_{\alpha_i} L(x, \alpha) \leq 0 \text{ for all } i \in [n], \tag{6.61}$$

Bounding the Error

we have

$$f(x) \geq f(\bar{x}) \geq f(x) + \sum_{i=1}^{m} \alpha_i c_i(x). \tag{6.62}$$

Strictly speaking, we only need differentiability of f and c_i at $\bar{x}$. However, since $\bar{x}$ is only known *after* the optimization problem has been solved, this is not a very useful condition.

Proof The first part of (6.62) follows from the fact that $x \in X$, so that x satisfies the constraints. Next note that $L(\bar{x}, \bar{\alpha}) = f(\bar{x})$ where $(\bar{x}, \bar{\alpha})$ denotes the saddle point of L. For the second part note that due to the saddle point condition (6.40), we have for any α with $\alpha_i \geq 0$,

$$f(\bar{x}) = L(\bar{x}, \bar{\alpha}) \geq L(\bar{x}, \alpha) \geq \inf_{x' \in X} L(x', \alpha). \tag{6.63}$$

The function $L(x', \alpha)$ is convex in x' since both f' and the constraints c_i are convex and all $\alpha_i \geq 0$. Therefore (6.60) implies that x minimizes $L(x', \alpha)$. This proves the second part of (6.63), which in turn proves the second inequality of (6.62). ∎

Hence, no matter what algorithm we are using in order to solve (6.37), we may always use (6.62) to assess the proximity of the current set of parameters to the solution. Clearly, the relative size of $\sum_{i=1}^{n} \alpha_i c_i(x)$ provides a useful stopping criterion for convex optimization algorithms.

Finally, another concept that is useful when dealing with optimization problems is that of *duality*. This means that for the *primal* minimization problem considered so far, which is expressed in terms of x, we can find a *dual* maximization problem in terms of α by computing the saddle point of the Lagrangian $L(x, \alpha)$, and eliminating the primal variables x. We thus obtain the following dual maximization problem from (6.37):

$$\begin{aligned} &\text{maximize} \quad L(x, \alpha) = f(x) + \sum_{i=1}^{n} \alpha_i c_i(x), \\ &\text{where} \quad (x, \alpha) \in Y := \left\{ (x, \alpha) \;\middle|\; \begin{array}{l} x \in X, \alpha_i \geq 0 \text{ for all } i \in [n] \\ \text{and } \partial_x L(x, \alpha) = 0 \end{array} \right\}. \end{aligned} \tag{6.64}$$

We state without proof a theorem guaranteeing the existence of a solution to (6.64).

Existence of Dual Solution

Theorem 6.28 (Wolfe [607]) *Recall the definition of* X *(6.45) and of the optimization problem (6.37). Under the assumptions that* $\mathcal{X}$ *is an open set,* X *satisfies one of the constraint qualifications of Lemma 6.23, and* f, c_i *are all convex and differentiable, there exists an* $\bar{\alpha} \in \mathbb{R}^n$ *such that* $(\bar{x}, \bar{\alpha})$ *solves the dual optimization problem (6.64) and in addition* $L(\bar{x}, \bar{\alpha}) = f(\bar{x})$.

In order to prove Theorem 6.28 we first have to show that some $(\bar{x}, \bar{\alpha})$ exists satisfying the KKT conditions, and then use the fact that the KKT-Gap at the saddle point vanishes.

6.3.3 Linear and Quadratic Programs

Primal Linear Program

Let us analyze the notions of primal and dual objective functions in more detail by looking at linear and quadratic programs. We begin with a simple linear setting.[6]

$$\begin{aligned} \underset{x}{\text{minimize}} \quad & c^\top x \\ \text{subject to} \quad & Ax + d \leq 0 \end{aligned} \tag{6.65}$$

where $c, x \in \mathbb{R}^m$, $d \in \mathbb{R}^n$ and $A \in R^{n \times m}$, and where $Ax + d \leq 0$ is a shorthand for $\sum_{j=1}^m A_{ij} x_j + d_i \leq 0$ for all $i \in [n]$.

Unbounded and Infeasible Problems

It is far from clear that (6.65) always has a solution, or indeed a minimum. For instance, the set of x satisfying $Ax + d \leq 0$ might be empty, or it might contain rays going to infinity in directions where $c^\top x$ keeps increasing. Before we deal with this issue in more detail, let us compute the sufficient KKT conditions for optimality, and the dual of (6.65). We may use (6.26) since (6.65) is clearly differentiable and convex. In particular we obtain:

Theorem 6.29 (KKT Conditions for Linear Programs) *A sufficient condition for a solution to the linear program (6.65) to exist is that the following four conditions are satisfied for some $(x, \alpha) \in \mathbb{R}^{m+n}$ where $\alpha \geq 0$:*

$$\partial_x L(x, \alpha) = \partial_x \left[c^\top x + \alpha^\top (Ax + d) \right] = A^\top \alpha + c = 0, \tag{6.66}$$

$$\partial_\alpha L(x, \alpha) = Ax + d \leq 0, \tag{6.67}$$

$$\alpha^\top (Ax + d) = 0, \tag{6.68}$$

$$\alpha \geq 0. \tag{6.69}$$

Then the minimum is given by $c^\top x$.

Note that, depending on the choice of A and d, there may not always exist an x such that $Ax + d \leq 0$, in which case the constraint does not satisfy the conditions of Lemma 6.23. In this situation, no solution exists for (6.65). If a feasible x exists, however, then (projections onto lower dimensional subspaces aside) the constraint qualifications are satisfied on the feasible set, and the conditions above are necessary. See [334, 345, 555] for details.

6. Note that we encounter a small clash of notation in (6.65), since c is used as a symbol for the loss function in the remainder of the book. This inconvenience is outweighed, however, by the advantage of consistency with the standard literature (e.g., [345, 45, 555]) on optimization. The latter will allow the reader to read up on the subject without any need for cumbersome notational changes.

Dual Linear Program

Next we may compute Wolfe's dual optimization problem by substituting (6.66) into $L(x, \alpha)$. Consequently, the primal variables x vanish, and we obtain a maximization problem in terms of α only:

$$\begin{aligned} &\text{maximize} && d^\top \alpha, \\ &\text{subject to} && A^\top \alpha + c = 0 \text{ and } \alpha \geq 0. \end{aligned} \tag{6.70}$$

Note that the number of variables and constraints has changed: we started with m variables and n constraints. Now we have n variables together with m equality constraints and n inequality constraints. While it is not yet completely obvious in the linear case, dualization may render optimization problems more amenable to numerical solution (the contrary may be true as well, though).

Primal Solution $\Leftrightarrow$ Dual Solution

What happens if a solution $\bar{x}$ to the primal problem (6.65) exists? In this case we know (since the KKT conditions of Theorem 6.29 are necessary and sufficient) that there must be an $\bar{\alpha}$ solving the dual problem, since $L(x, \alpha)$ has a saddle point at $(\bar{x}, \bar{\alpha})$.

If no feasible point of the primal problem exists, there must exist, by (a small modification of) Lemma 6.23, some $\alpha \in \mathbb{R}^n$ with $\alpha \geq 0$ and at least one $\alpha_i > 0$ such that $\alpha^\top (Ax + d) > 0$ for all x. This means that for all x, the Lagrangian $L(x, \alpha)$ is unbounded from above, since we can make $\alpha^\top (Ax + d)$ arbitrarily large. Hence the dual optimization problem is unbounded. Using analogous reasoning, if the primal problem is unbounded, the dual problem is infeasible.

Let us see what happens if we dualize (6.70) one more time. First we need more Lagrange multipliers, since we have two sets of constraints. The equality constraints can be taken care of by an unbounded variable x' (see Theorem 6.22 for how to deal with equalities). For the inequalities $\alpha \geq 0$, we introduce a second Lagrange multiplier $y \in \mathbb{R}^n$. After some calculations and resubstitution into the corresponding Lagrangian, we get

$$\begin{aligned} &\text{maximize} && c^\top x', \\ &\text{subject to} && Ax' + d + y = 0 \text{ and } y \geq 0. \end{aligned} \tag{6.71}$$

Dual Dual Linear Program $\rightarrow$ Primal

We can remove $y \geq 0$ from the set of variables by transforming $Ax' + d + y$ into $Ax + d \leq 0$; thus we recover the primal optimization problem (6.65).[7]

The following theorem gives an overview of the transformations and relations between primal and dual problems (see also Table 6.2). Although we only derived these relations for linear programs, they also hold for other convex differentiable settings [45].

Theorem 6.30 (Trichotomy) *For linear and convex quadratic programs exactly one of*

7. This finding is useful if we have to dualize twice in some optimization settings (see Chapter 10), since then we will be able to recover some of the primal variables without further calculations if the optimization algorithm provides us with both primal and dual variables.

Table 6.2 Connections between primal and dual linear and convex quadratic programs.

Primal Optimization Problem (in x)	Dual Optimization Problem (in α)
solution exists	solution exists and extrema are equal
no solution exists	maximization problem has unbounded objective from above or is infeasible
minimization problem has unbounded objective from below or is infeasible	no solution exists
inequality constraint	inequality constraint
equality constraint	free variable
free variable	equality constraint

the following three alternatives must hold:

1. *Both feasible regions are empty.*
2. *Exactly one feasible region is empty, in which case the objective function of the other problem is unbounded in the direction of optimization.*
3. *Both feasible regions are nonempty, in which case both problems have solutions and their extrema are equal.*

Primal Quadratic Program

We conclude this section by stating primal and dual optimization problems, and the sufficient KKT conditions for convex quadratic optimization problems. To keep matters simple we only consider the following type of optimization problem (other problems can be rewritten in the same form; see Problem 6.11 for details):

$$\begin{aligned} &\underset{x}{\text{minimize}} && \tfrac{1}{2}x^\top Kx + c^\top x, \\ &\text{subject to} && Ax + d \leq 0. \end{aligned} \tag{6.72}$$

Here K is a strictly positive definite matrix, $x, c \in \mathbb{R}^m$, $A \in \mathbb{R}^{n\times m}$, and $d \in \mathbb{R}^n$. Note that this is clearly a differentiable convex optimization problem. To introduce a Lagrangian we need corresponding multipliers $\alpha \in \mathbb{R}^n$ with $\alpha \geq 0$. We obtain

$$L(x,\alpha) = \frac{1}{2}x^\top Kx + c^\top x + \alpha^\top(Ax + d). \tag{6.73}$$

Next we may apply Theorem 6.26 to obtain the KKT conditions. They can be stated in analogy to (6.66)–(6.68) as

$$\partial_x L(x,\alpha) = \partial_x \left[c^\top x + \alpha^\top(Ax+d) + \frac{1}{2}x^\top Kx\right] = Kx + A^\top\alpha + c = 0, \tag{6.74}$$

$$\partial_\alpha L(x,\alpha) = Ax + d \leq 0, \tag{6.75}$$

$$\alpha^\top(Ax + d) = 0, \tag{6.76}$$

$$\alpha \geq 0. \tag{6.77}$$

In order to compute the dual of (6.72), we have to eliminate x from (6.73) and write it as a function of α. We obtain

$$L(x,\alpha) = -\frac{1}{2}x^\top Kx + \alpha^\top d \tag{6.78}$$

$$= -\frac{1}{2}\alpha^\top A^\top K^{-1}A\alpha + \left[d - c^\top K^{-1}A^\top\right]\alpha - \frac{1}{2}c^\top K^{-1}c. \tag{6.79}$$

In (6.78) we used (6.74) and (6.76) directly, whereas in order to eliminate x completely in (6.79) we solved (6.74) for $x = -K^{-1}(c + A^\top\alpha)$. Ignoring constant terms this leads to the dual quadratic optimization problem,

Dual Quadratic Program

$$\begin{array}{ll}\underset{\alpha}{\text{minimize}} & -\frac{1}{2}\alpha^\top A^\top K^{-1}A\alpha + \left[d - c^\top K^{-1}A^\top\right]\alpha, \\ \text{subject to} & \alpha \geq 0.\end{array} \tag{6.80}$$

The surprising fact about the dual problem (6.80) is that the constraints become significantly simpler than in the primal (6.72). Furthermore, if $n < m$, we also obtain a more compact representation of the quadratic term.

There is one aspect in which (6.80) differs from its linear counterpart (6.70): if we dualize (6.80) again, we do not recover (6.72) but rather a problem very similar in structure to (6.80). Dualizing (6.80) twice, however, we recover the dual itself (Problem 6.13 deals with this matter in more detail).

6.4 Interior Point Methods

Let us now have a look at simple, yet efficient optimization algorithms for constrained problems: interior point methods.

An interior point is a pair of variables (x, α) that satisfies both primal and dual constraints. As already mentioned before, finding a set of vectors $(\bar{x}, \bar{\alpha})$ that satisfy the KKT conditions is sufficient to obtain a solution in $\bar{x}$. Hence, all we have to do is devise an algorithm which solves (6.74)–(6.77), for instance, if we want to solve a quadratic program. We will focus on the quadratic case — the changes required for linear programs merely involve the removal of some variables, simplifying the equations. See Problem 6.14 and [555, 517] for details.

6.4.1 Sufficient Conditions for a Solution

We need a slight modification of (6.74)–(6.77) in order to achieve our goal: rather than the inequality (6.75), we are better off with an equality and a positivity constraint for an additional variable, i.e. we transform $Ax + d \leq 0$ into $Ax + d + \xi =$