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

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

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 $Ax + d + \xi =$

0, where $\xi \geq 0$. Hence we arrive at the following system of equations:

$$\begin{aligned} Kx + A^\top \alpha + c &= 0 \quad \text{(Dual Feasibility)}, \\ Ax + d + \xi &= 0 \quad \text{(Primal Feasibility)}, \\ \alpha^\top \xi &= 0, \\ \alpha, \xi &\geq 0. \end{aligned} \tag{6.81}$$

Optimality as Constraint Satisfaction

Let us analyze the equations in more detail. We have three sets of variables: x, α, ξ. To determine the latter, we have an equal number of equations plus the positivity constraints on α, ξ. While the first two equations are linear and thus amenable to solution, e.g., by matrix inversion, the third equality $\alpha^\top \xi = 0$ has a small defect: given one variable, say α, we cannot solve it for ξ or vice versa. Furthermore, the last two constraints are not very informative either.

We use a primal-dual path-following algorithm, as proposed in [556], to solve this problem. Rather than requiring $\alpha^\top \xi = 0$ we modify it to become $\alpha_i \xi_i = \mu > 0$ for all $i \in [n]$, solve (6.81) for a given μ, and decrease μ to 0 as we go. The advantage of this strategy is that we may use a Newton-type predictor corrector algorithm (see Section 6.2.5) to update the parameters x, α, ξ, which exhibits the fast convergence of a second order method.

6.4.2 Solving the Equations

For the moment, assume that we have suitable initial values of x, α, ξ, and μ with $\alpha, \xi > 0$. Linearization of the first three equations of (6.81), together with the modification $\alpha_i \xi_i = \mu$, yields (we expand x into $x + \Delta x$, etc.):

Linearized Constraints

$$\begin{aligned} K\Delta x + A^\top \Delta\alpha &= -Kx - A^\top \alpha - c &&=: \rho_p, \\ A\Delta x + \Delta\xi &= -Ax - d - \xi &&=: \rho_d, \\ \alpha_i^{-1}\xi_i \Delta\alpha_i + \Delta\xi_i &= \mu\alpha_i^{-1} - \xi_i - \alpha_i^{-1}\Delta\alpha_i\Delta\xi_i &&=: \rho_{\mathrm{KKT}i} \text{ for all } i \end{aligned} \tag{6.82}$$

Next we solve for $\Delta\xi_i$ to obtain what is commonly referred to as the *reduced* KKT system. For convenience we use $D := \operatorname{diag}(\alpha_1^{-1}\xi_1, \ldots, \alpha_n^{-1}\xi_n)$ as a shorthand;

$$\begin{bmatrix} K & A^\top \\ A & -D \end{bmatrix} \begin{bmatrix} \Delta x \\ \Delta\alpha \end{bmatrix} = \begin{bmatrix} \rho_p \\ \rho_d - \rho_{\mathrm{KKT}} \end{bmatrix}. \tag{6.83}$$

We apply a predictor-corrector method as in Section 6.2.5. The resulting matrix of the linear system in (6.83) is indefinite but of full rank, and we can solve (6.83) for $(\Delta x_{\mathrm{Pred}}, \Delta\alpha_{\mathrm{Pred}})$ by explicitly pivoting for individual entries (for instance, solve for Δx first and then substitute the result in to the second equality to obtain $\Delta\alpha$).

This gives us the *predictor* part of the solution. Next we have to correct for the linearization, which is conveniently achieved by updating ρ_{KKT} and solving (6.83) again to obtain the *corrector* values $(\Delta x_{\mathrm{Corr}}, \Delta\alpha_{\mathrm{Corr}})$. The value of $\Delta\xi$ is then obtained from (6.82).

Update in x, α

Next, we have to make sure that the updates in α, ξ do not cause the estimates to violate their positivity constraints. This is done by shrinking the length of $(\Delta x, \Delta\alpha, \Delta\xi)$ by some factor $\lambda \geq 0$, such that

$$\min\left(\frac{\alpha_1 + \lambda\Delta\alpha_1}{\alpha_1}, \ldots, \frac{\alpha_n + \lambda\Delta\alpha_n}{\alpha_n}, \frac{\xi_1 + \lambda\Delta\xi_1}{\xi_1}, \ldots, \frac{\xi_n + \lambda\Delta\xi_n}{\xi_n}\right) \geq \epsilon. \tag{6.84}$$

Of course, only the negative Δ terms pose a problem, since they lead the parameter values closer to 0, which may lead them into conflict with the positivity constraints. Typically [556, 502], we choose $\epsilon = 0.05$. In other words, the solution will not approach the boundaries in α, ξ by more than 95%. See Problem 6.15 for a formula to compute λ.

6.4.3 Updating μ

Next we have to update μ. Here we face the following dilemma: if we decrease μ too quickly, we will get bad convergence of our second order method, since the solution to the problem (which depends on the value of μ) moves too quickly away from our current set of parameters (x, α, ξ). On the other hand, we do not want to spend too much time solving an *approximation* of the unrelaxed ($\mu = 0$) KKT conditions *exactly*. A good indication is how much the positivity constraints would be violated by the current update. Vanderbei [556] proposes the following update of μ:

Tightening the KKT Conditions

$$\mu = \frac{\alpha^\top \xi}{n}\left(\frac{1-\lambda}{10+\lambda}\right)^2. \tag{6.85}$$

The first term gives the average value of satisfaction of the condition $\alpha_i \xi_i = \mu$ after an update step. The second term allows us to decrease μ rapidly if good progress was made (small $(1-\lambda)^2$). Experimental evidence shows that it pays to be slightly more conservative, and to use the *predictor* estimates of α, ξ for (6.85) rather than the corresponding corrector terms.[8] This imposes little overhead for the implementation.

6.4.4 Initial Conditions and Stopping Criterion

To provide a complete algorithm, we have to consider two more things: a stopping criterion and a suitable start value. For the latter, we simply solve a regularized version of the initial reduced KKT system (6.83). This means that we replace K by $K + \mathbf{1}$, use (x, α) in place of $\Delta x, \Delta\alpha$, and replace D by the identity matrix. Moreover, ρ_p and ρ_d are set to the values they would have if all variables had been set to 0 before, and finally ρ_{KKT} is set to 0. In other words, we obtain an initial guess of

Regularized KKT System

[8] In practice it is often useful to replace $(1-\lambda)$ by $(1+\epsilon-\lambda)$ for some small $\epsilon > 0$, in order to avoid $\mu = 0$.

(x, α, ξ) by solving

$$\begin{bmatrix} K + \mathbf{1} & A^\top \\ A & -\mathbf{1} \end{bmatrix} \begin{bmatrix} x \\ \alpha \end{bmatrix} = \begin{bmatrix} -c \\ -d \end{bmatrix}, \tag{6.86}$$

and $\xi = -Ax - d$. Since we have to ensure positivity of α, ξ, we simply replace

$$\alpha_i = \max(\alpha_i, 1) \text{ and } \xi_i = \max(\xi_i, 1). \tag{6.87}$$

This heuristic solves the problem of a suitable initial condition.

Regarding the stopping criterion, we recall Theorem 6.27, and in particular (6.62). Rather than obtaining bounds on the precision of *parameters*, we want to make sure that $f(x)$ is close to its optimal value $f(\bar{x})$. From (6.64) we know, provided the feasibility constraints are all satisfied, that the value of the dual objective function is given by $f(x) + \sum_{i=1}^n \alpha_i c_i(x)$. We may use the latter to bound the *relative* size of the gap between primal and dual objective function by

$$\operatorname{Gap}(x, \alpha) = \frac{2\left| f(x) - \left(f(x) + \sum\limits_{i=1}^n \alpha_i c_i(x) \right) \right|}{|f(x)| + \left| \left(f(x) + \sum\limits_{i=1}^n \alpha_i c_i(x) \right) \right|} \leq \frac{-\sum\limits_{i=1}^n \alpha_i c_i(x)}{\left| f(x) + \frac{1}{2} \sum\limits_{i=1}^n \alpha_i c_i(x) \right|}. \tag{6.88}$$

For the special case where $f(x) = \frac{1}{2}x^\top K x + c^\top x$ as in (6.72), we know by virtue of (6.73) that the size of the feasibility gap is given by $\alpha^\top \xi$, and therefore

$$\operatorname{Gap}(x, \alpha) = \frac{\alpha^\top \xi}{\left| \frac{1}{2} x^\top K x + c^\top x + \frac{1}{2} \alpha^\top \xi \right|}. \tag{6.89}$$

Number of Significant Figures

In practice, a small number is usually added to the denominator of (6.89) in order to avoid divisions by 0 in the first iteration. The quality of the solution is typically measured on a logarithmic scale by $-\log_{10} \operatorname{Gap}(x, \alpha)$, the number of significant figures.[9] We will come back to specific versions of such interior point algorithms in Chapter 10, and show how Support Vector Regression and Classification problems can be solved with them.

Primal-Dual path following methods are certainly not the only algorithms that can be employed for minimizing constrained quadratic problems. Other variants, for instance, are Barrier Methods [282, 45, 557], which minimize the unconstrained problem

$$f(x) + \mu \sum_{i=1}^n f \ln(-c_i(x)) \text{ for } \mu > 0. \tag{6.90}$$

Active set methods have also been used with success in machine learning [369, 284]. These select subsets of variables x for which the constraints c_i are not ac-

[9] Interior point codes are very precise. They usually achieve up to 8 significant figures, whereas iterative approximation methods do not normally exceed more than 3 significant figures on large optimization problems.

tive, i.e., where the we have a strict inequality, and solve the resulting restricted quadratic program, for instance by conjugate gradient descent. We will encounter subset selection methods in Chapter 10.

6.5 Maximum Search Problems

Approximations

In several cases the task of finding an optimal function for estimation purposes means finding the best element from a finite set, or sometimes finding an optimal subset from a finite set of elements. These are discrete (sometimes combinatorial) optimization problems which are not so easily amenable to the techniques presented in the previous two sections. Furthermore, many commonly encountered problems are computationally expensive if solved exactly. Instead, by using probabilistic methods, it is possible to find *almost* optimal approximate solutions. These probabilistic methods are the topic of the present section.

6.5.1 Random Subset Selection

Consider the following problem: given a set of m functions, say $M := \{f_1, \ldots, f_m\}$, and some criterion $Q[f]$, find the function $\hat{f}$ that maximizes $Q[f]$. More formally,

$$\hat{f} := \operatorname*{argmax}_{f \in M} Q[f]. \tag{6.91}$$

Clearly, unless we have additional knowledge about the values $Q[f_i]$, we have to compute all terms $Q[f_i]$ if we want to solve (6.91) exactly. This will cost $O(m)$ operations. If m is large, which is often the case in practical applications, this operation is too expensive. In sparse greedy approximation problems (Section 10.2) or in Kernel Feature Analysis (Section 14.4), m can easily be of the order of 10^5 or larger (here, m is the number of training patterns). Hence we have to look for cheaper *approximate* solutions.

The key idea is to pick a random subset $M' \subset M$ that is sufficiently large, and take the maximum over M' as an approximation of the maximum over M. Provided the distribution of the values of $Q[f_i]$ is "well behaved", i.e., there exists not a small fraction of $Q[f_i]$ whose values are significantly smaller or larger than the average, we will obtain a solution that is close to the optimum with high probability. To formalize these ideas, we need the following result.

Lemma 6.31 (Maximum of Random Variables) *Denote by ξ, ξ' two independent random variables on $\mathbb{R}$ with corresponding distributions $\mathrm{P}_\xi, \mathrm{P}_{\xi'}$ and distribution functions $F_\xi, F_{\xi'}$. Then the random variable $\bar{\xi} := \max(\xi, \xi')$ has the distribution function $F_{\bar{\xi}} = F_\xi F_{\xi'}$.*

Proof Note that for a random variable, the distribution function $F(\xi_0)$ is given by