CHAPTER 19

The KKT System

The most time-consuming aspect of each iteration of the path-following method is solving the system of equations that defines the step direction vectors Δx, Δy, Δw, and Δz:

$$A\Delta x + \Delta w = \rho \tag{19.1}$$

$$A^T\Delta y - \Delta z = \sigma \tag{19.2}$$

$$Z\Delta x + X\Delta z = \mu e - XZe \tag{19.3}$$

$$W\Delta y + Y\Delta w = \mu e - YWe. \tag{19.4}$$

After minor manipulation, these equations can be written in block matrix form as follows:

$$\left[\begin{array}{cc|cc} -XZ^{-1} & & -I & \\ & & A & I \\ \hline -I & A^T & & \\ & I & & YW^{-1} \end{array}\right] \left[\begin{array}{c} \Delta z \\ \Delta y \\ \hline \Delta x \\ \Delta w \end{array}\right] = \left[\begin{array}{c} -\mu Z^{-1}e + x \\ \rho \\ \hline \sigma \\ \mu W^{-1}e - y \end{array}\right]. \tag{19.5}$$

This system is called the *Karush–Kuhn–Tucker system*, or simply the KKT system. It is a symmetric linear system of $2n + 2m$ equations in $2n + 2m$ unknowns. One could, of course, perform a factorization of this large system and then follow that with a forward and backward substitution to solve the system. However, it is better to do part of this calculation "by hand" first and only use a factorization routine to help solve a smaller system. There are two stages of reductions that one could apply. After the first stage, the remaining system is called the reduced KKT system, and after the second stage it is called the system of normal equations. We shall discuss these two systems in the next two sections.

1. The Reduced KKT System

Equations (19.3) and (19.4) are trivial (in the sense that they only involve diagonal matrices), and so it seems sensible to eliminate them right from the start. To preserve the symmetry that we saw in (19.5), we should solve them for Δz and Δw, respectively:

$$\Delta z = X^{-1}(\mu e - XZe - Z\Delta x)$$
$$\Delta w = Y^{-1}(\mu e - YWe - W\Delta y).$$

R.J. Vanderbei, *Linear Programming*, International Series in Operations Research & Management Science 196, DOI 10.1007/978-1-4614-7630-6_19,

Substituting these formulas into (19.1) and (19.2), we get the so-called *reduced KKT system*:

$$A\Delta x - Y^{-1}W\Delta y = \rho - \mu Y^{-1}e + w \tag{19.6}$$

$$A^T\Delta y + X^{-1}Z\Delta x = \sigma + \mu X^{-1}e - z. \tag{19.7}$$

Substituting in the definitions of ρ and σ and writing the system in matrix notation, we get

$$\begin{bmatrix} -Y^{-1}W & A \\ A^T & X^{-1}Z \end{bmatrix} \begin{bmatrix} \Delta y \\ \Delta x \end{bmatrix} = \begin{bmatrix} b - Ax - \mu Y^{-1}e \\ c - A^T y + \mu X^{-1}e \end{bmatrix}.$$

Note that the reduced KKT matrix is again a symmetric matrix. Also, the right-hand side displays symmetry between the primal and the dual. To reduce the system any further, one needs to break the symmetry that we have carefully preserved up to this point. Nonetheless, we forge ahead.

2. The Normal Equations

For the second stage of reduction, there are two choices: we could either (1) solve (19.6) for Δy and eliminate it from (19.7) or (2) solve (19.7) for Δx and eliminate it from (19.6). For the moment, let us assume that we follow the latter approach. In this case, we get from (19.7) that

$$\Delta x = XZ^{-1}(c - A^T y + \mu X^{-1}e - A^T\Delta y), \tag{19.8}$$

which we use to eliminate Δx from (19.6) to get

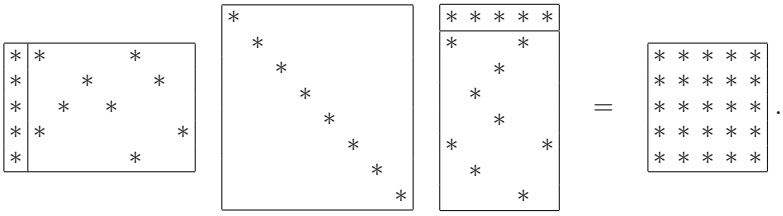
$$\begin{aligned} -(Y^{-1}W + AXZ^{-1}A^T)\Delta y = {} & b - Ax - \mu Y^{-1}e \\ & - AXZ^{-1}(c - A^T y + \mu X^{-1}e). \end{aligned} \tag{19.9}$$

This last system is a system of m equations in m unknowns. It is called the *system of normal equations in primal form*. It is a system of equations involving the matrix $Y^{-1}W + AXZ^{-1}A^T$. The $Y^{-1}W$ term is simply a diagonal matrix, and so the real meat of this matrix is contained in the $AXZ^{-1}A^T$ term.

Given that A is sparse (which is generally the case in real-world linear programs), one would expect the matrix $AXZ^{-1}A^T$ to be likewise sparse. However, we need to investigate the sparsity of $AXZ^{-1}A^T$ (or lack thereof) more closely. Note that the (i,j)th element of this matrix is given by

$$(AXZ^{-1}A^T)_{ij} = \sum_{k=1}^{n} a_{ik}\frac{x_k}{z_k}a_{jk}.$$

That is, the (i,j)th element is simply a weighted inner product of rows i and j of the A matrix. If these rows have disjoint nonzero patterns, then this inner product is guaranteed to be zero, but otherwise it must be treated as a potential nonzero. This is bad news if A is generally sparse but has, say, one dense column:



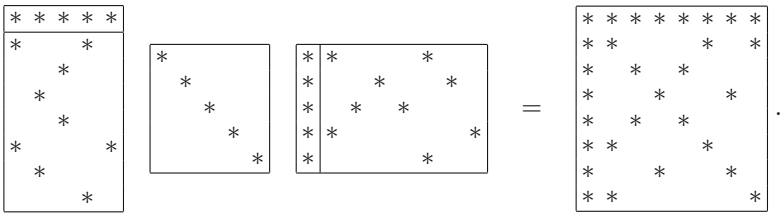
But don't forget that we didn't have to go the primal normal equations route. Instead, we could have chosen the other alternative of solving (19.6) for Δy,

$$\Delta y = -YW^{-1}(b - Ax - \mu Y^{-1}e - A\Delta x),$$

and eliminating it from (19.7):

$$(A^T YW^{-1}A + X^{-1}Z)\Delta x = c - A^T y + \mu X^{-1}e + A^T YW^{-1}(b - Ax - \mu Y^{-1}e). \tag{19.10}$$

The system defined by (19.10) is a system of n equations in n unknowns. It is called the system of *normal equations in dual form*. Note that dense columns do not pose a problem for these equations. Indeed, for the example given above, we now get



While this system is larger than the one before, it is also sparse, and sparsity almost always is more important than matrix dimensions. In this example, the dense matrix associated with the primal normal equations requires 65 arithmetic operations to factor, whereas the larger, sparser matrix associated with the dual normal equations requires just 60. This is a small difference, but these are small matrices. As the matrices involved get large, factoring a dense matrix takes on the order of n^3 operations, whereas a very sparse matrix might take only on the order of n operations. Clearly, as n gets large, the difference between these becomes quite significant.

It would be great if we could say that it is always best to solve the primal normal equations or the dual normal equations. But as we've just seen, dense columns in A are bad for the primal normal equations and, of course, it follows that dense rows are bad for the dual normal equations. Even worse, some problems have constraint matrices A that are overall very sparse but contain some dense rows and some dense columns. Such problems are sure to run into trouble with either sets of normal equations. For these reasons, it is best to factor the matrix in the reduced KKT system directly. Then it is possible to find pivot orders that circumvent the difficulties posed by both dense columns and dense rows.

3. Step Direction Decomposition

In the next chapter, we shall discuss factorization techniques for symmetric matrices (along with other implementation issues). However, before we embark on that discussion, we end this chapter by taking a closer look at the formulas for the step direction vectors. To be specific, let us look at Δx. From the primal normal equations (19.9), we can solve for Δy and then substitute the solution into (19.8) to get an explicit formula for Δx:

$$\begin{aligned}\Delta x = &\left(D^2 - D^2A^T(E^{-2} + AD^2A^T)^{-1}AD^2\right)(c - A^Ty + \mu X^{-1}e)\\ &+ D^2A^T(E^{-2} + AD^2A^T)^{-1}(b - Ax - \mu Y^{-1}e),\end{aligned} \tag{19.11}$$

where we have denoted by D the positive diagonal matrix defined by

$$D^2 = XZ^{-1}$$

and we have denoted by E the positive diagonal matrix defined by

$$E^2 = YW^{-1}$$

(defining these matrices by their squares is possible, since the squares have positive diagonal entries). However, using the dual normal equations, we get

$$\begin{aligned}\Delta x = &\left(A^TE^2A + D^{-2}\right)^{-1}(c - A^Ty + \mu X^{-1}e)\\ &+ \left(A^TE^2A + D^{-2}\right)^{-1}A^TE^2(b - Ax - \mu Y^{-1}e).\end{aligned} \tag{19.12}$$

These two expressions for Δx look entirely different, but they must be the same, since we know that Δx is uniquely defined by the reduced KKT system. They are indeed the same, as can be shown directly by establishing a certain matrix identity. This is the subject of Exercise 19.1. There are a surprising number of published research papers on interior-point methods that present supposedly new algorithms that are in fact identical to existing methods. These papers get published because the equivalence is not immediately obvious (such as the one we just established).

We can gain further insight into the path-following method by looking more closely at the primal step direction vector. Formula (19.11) can be rearranged as follows:

$$\begin{aligned}\Delta x = &\left(D^2 - D^2A^T(E^{-2} + AD^2A^T)^{-1}AD^2\right)c\\ &+ \mu\left(D^2 - D^2A^T(E^{-2} + AD^2A^T)^{-1}AD^2\right)X^{-1}e\\ &- \mu D^2A^T(E^{-2} + AD^2A^T)^{-1}Y^{-1}e\\ &+ D^2A^T(E^{-2} + AD^2A^T)^{-1}(b - Ax)\\ &- D^2A^T\left(I - (E^{-2} + AD^2A^T)^{-1}AD^2A^T\right)y.\end{aligned}$$

For comparison purposes down the road, we prefer to write the $Y^{-1}e$ that appears in the second term containing μ as $E^{-2}W^{-1}e$. Also, using the result of Exercise 19.2, we can rewrite the bulk of the last line as follows:

$$\begin{aligned}\left(I - (E^{-2} + AD^2A^T)^{-1}AD^2A^T\right)y &= (E^{-2} + AD^2A^T)^{-1}E^{-2}y\\ &= (E^{-2} + AD^2A^T)^{-1}w.\end{aligned}$$

Putting this all together, we get

$$\begin{aligned}\Delta x &= \left(D^2 - D^2A^T(E^{-2} + AD^2A^T)^{-1}AD^2\right)c \\ &\quad + \mu\left(D^2 - D^2A^T(E^{-2} + AD^2A^T)^{-1}AD^2\right)X^{-1}e \\ &\quad - \mu D^2A^T(E^{-2} + AD^2A^T)^{-1}E^{-2}W^{-1}e \\ &\quad + D^2A^T(E^{-2} + AD^2A^T)^{-1}\rho \\ &= \Delta x_{\text{OPT}} + \mu\Delta x_{\text{CTR}} + \Delta x_{\text{FEAS}},\end{aligned}$$

where

$$\begin{aligned}\Delta x_{\text{OPT}} &= \left(D^2 - D^2A^T(E^{-2} + AD^2A^T)^{-1}AD^2\right)c, \\ \Delta x_{\text{CTR}} &= \left(D^2 - D^2A^T(E^{-2} + AD^2A^T)^{-1}AD^2\right)X^{-1}e \\ &\quad - D^2A^T(E^{-2} + AD^2A^T)^{-1}E^{-2}W^{-1}e,\end{aligned}$$

and

$$\Delta x_{\text{FEAS}} = D^2A^T(E^{-2} + AD^2A^T)^{-1}\rho.$$

In Chapter 21, we shall show that these components of Δx have important connections to the step directions that arise in a related interior-point method called the affine-scaling method. For now, we simply establish some of their properties as they relate to the path-following method. Our first result is that Δx_{OPT} is an ascent direction.

THEOREM 19.1. $c^T\Delta x_{\text{OPT}} \geq 0$.

PROOF. We use the result of Exercise 19.1 (with the roles of E and D switched) to see that

$$\Delta x_{\text{OPT}} = (A^TE^2A + D^{-2})^{-1}c.$$

Hence,

$$c^T\Delta x_{\text{OPT}} = c^T(A^TE^2A + D^{-2})^{-1}c.$$

We claim that the right-hand side is obviously nonnegative, since the matrix sandwiched between c and its transpose is positive semidefinite.[1] Indeed, the claim follows from the definition of positive semidefiniteness: a matrix B is *positive semidefinite* if $\xi^TB\xi \geq 0$ for all vectors ξ. To see that the matrix in question is in fact positive semidefinite, we first note that A^TE^2A and D^{-2} are positive semidefinite:

$$\xi^TA^TE^2A\xi = \|EA\xi\|^2 \geq 0 \qquad \text{and} \qquad \xi^TD^{-2}\xi = \|D^{-1}\xi\|^2 \geq 0.$$

Then we show that the sum of two positive semidefinite matrices is positive semidefinite and finally that the inverse of a symmetric positive semidefinite matrix is positive semidefinite. To verify closure under summation, suppose that $B^{(1)}$ and $B^{(2)}$ are positive semidefinite, and then compute

$$\xi^T(B^{(1)} + B^{(2)})\xi = \xi^TB^{(1)}\xi + \xi^TB^{(2)}\xi \geq 0.$$

To verify closure under forming inverses of symmetric positive semidefinite matrices, suppose that B is symmetric and positive semidefinite. Then

[1] In fact, it's positive definite, but we don't need this stronger property here.

$$\xi^T B^{-1}\xi = \xi^T B^{-1} B B^{-1}\xi = (B^{-1}\xi)^T B (B^{-1}\xi) \geq 0,$$

where the inequality follows from the fact that B is positive semidefinite and $B^{-1}\xi$ is simply any old vector. This completes the proof. □

The theorem just proved justifies our referring to Δx_{OPT} as a *step-toward-optimality* direction. We next show that Δx_{FEAS} is in fact a *step-toward-feasibility*.

In Exercise 19.3, you are asked to find the formulas for the primal slack vector's step directions, Δw_{OPT}, Δw_{CTR}, and Δw_{FEAS}. It is easy to verify from these formulas that the pairs $(\Delta x_{\text{OPT}}, \Delta w_{\text{OPT}})$ and $(\Delta x_{\text{CTR}}, \Delta w_{\text{CTR}})$ preserve the current level of infeasibility. That is,

$$A\Delta x_{\text{OPT}} + \Delta w_{\text{OPT}} = 0$$

and

$$A\Delta x_{\text{CTR}} + \Delta w_{\text{CTR}} = 0.$$

Hence, only the "feasibility" directions can improve the degree of feasibility. Indeed, it is easy to check that

$$A\Delta x_{\text{FEAS}} + \Delta w_{\text{FEAS}} = \rho.$$

Finally, we consider Δx_{CTR}. If the objective function were zero (i.e., $c = 0$) and if the current point were feasible, then steps toward optimality and feasibility would vanish and we would be left with just Δx_{CTR}. Since our step directions were derived in an effort to move toward a point on the central path parametrized by μ, we now see that Δx_{CTR} plays the role of a *step-toward-centrality*.

Exercises

19.1 *Sherman–Morrison–Woodbury Formula.* Assuming that all the inverses below exist, show that the following identity is true:

$$(E^{-1} + ADA^T)^{-1} = E - EA(A^T EA + D^{-1})^{-1} A^T E.$$

Use this identity to verify directly the equivalence of the expressions given for Δx in (19.11) and (19.12).

19.2 Assuming that all the inverses exist, show that the following identity holds:

$$I - (E + ADA^T)^{-1} ADA^T = (E + ADA^T)^{-1} E.$$

19.3 Show that

$$\Delta w = \Delta w_{\text{OPT}} + \mu \Delta w_{\text{CTR}} + \Delta w_{\text{FEAS}},$$

where

$$\begin{aligned}
\Delta w_{\text{OPT}} &= -A\left(D^2 - D^2 A^T (E^{-2} + AD^2 A^T)^{-1} AD^2\right) c, \\
\Delta w_{\text{CTR}} &= -A\left(D^2 - D^2 A^T (E^{-2} + AD^2 A^T)^{-1} AD^2\right) X^{-1} e \\
&\quad + AD^2 A^T (E^{-2} + AD^2 A^T)^{-1} E^{-2} W^{-1} e,
\end{aligned}$$

and

$$\Delta w_{\text{FEAS}} = \rho - AD^2 A^T (E^{-2} + AD^2 A^T)^{-1} \rho.$$

Notes

The KKT system for general inequality constrained optimization problems was derived by Kuhn and Tucker (1951). It was later discovered that W. Karush had proven the same result in his 1939 master's thesis at the University of Chicago (Karush 1939). John (1948) was also an early contributor to inequality-constrained optimization. Kuhn's survey paper (Kuhn 1976) gives a historical account of the development of the subject.