PREPRINT ANL/MCS-P600-0596, MAY, 1996, MATHEMATICS AND COMPUTER SCIENCE DIVISION, ARGONNE NATIONAL LABORATORY, ARGONNE, IL 60439, USA.

MODIFIED CHOLESKY FACTORIZATIONS IN INTERIOR-POINT ALGORITHMS FOR LINEAR PROGRAMMING

STEPHEN WRIGHT*

Abstract. We investigate a modified Cholesky algorithm similar to those used in current interior-point codes for linear programming. Cholesky-based interior-point codes are popular for three reasons: their implementation requires only minimal changes to standard sparse Cholesky codes (allowing us to take full advantage of software written by specialists in that area); they tend to be more efficient than competing approaches that use different factorizations; and they perform robustly on most practical problems, yielding good interior-point steps even when the coefficient matrix is ill conditioned. We explain the surprisingly good performance of the Cholesky-based approach by using analytical tools from matrix perturbation theory and error analysis, illustrating our results with computational experiments. Finally, we point out the limitations of this approach.

Key words. Interior-point algorithms and software, Cholesky factorization, Matrix perturbations, Error analysis.

1. Introduction.

Most interior-point codes for linear programming share a common feature: their major computational operation—solution of a large linear system of equations—is performed by a direct sparse Cholesky algorithm. In this algorithm, row and column orderings are determined a priori by well-known heuristics (minimum degree and enhancements, minimum local fill, nested dissection) that are based solely on the sparsity pattern and not on the numerical values of the nonzero elements. The ordering phase is followed by a symbolic factorization phase, in which the nonzero structure of the Cholesky factor is determined and storage is allocated. Finally, a numerical factorization phase fills in the numerical values of the lower triangular Cholesky factor. In interior-point codes, the first two phases usually are performed just once, during either the first interior-point iteration or computation of a starting point.

In the interior-point context, the unadorned Cholesky algorithm can run into difficulties because of extreme ill conditioning. Some of the diagonal pivots encountered during the numerical factorization phase can be zero or negative, causing the standard Cholesky procedure to break down. Instead of crashing, most codes apply a "patch" to the algorithm to handle such pivots. The offending pivot element is sometimes replaced by a huge number, as in LIPSOL [17] or PCx [1]. In other codes such as IPMOS [16], the pivot is replaced by a moderate number, but the corresponding right-hand side element is set to zero, as are the off-diagonal elements in the corresponding column of the Cholesky factor. The first practical interior-point code, OB1 [6], explicitly zeroes the components of the solution vector that correspond to small pivots. All these strategies are essentially equivalent to the algorithm we describe in this paper. To date, there has been little investigation of them from a numerical analysis viewpoint.

The "patches" described above have the advantage that they can be implemented by changing just a few lines in general sparse Cholesky codes. It is therefore possible to take advantage of the long-term development effort that has gone into designing such codes and their underlying algorithms. The recent codes LIPSOL [17] and PCx

* Mathematics and Computer Science Division, Argonne National Laboratory, 9700 South Cass Avenue, Argonne, IL 60439. This work was supported by the Mathematical, Information, and Computational Sciences Division subprogram of the Office of Computational and Technology Research, U.S. Department of Energy, under Contract W-31-109-Eng-38.

[1] make explicit use of the freely available sparse Cholesky code of Ng and Peyton [8]. Other codes either modify the well-known SPARSPAK routines of George and Liu [3] or include customized linear algebra routines that implement well established algorithmic ideas. (At least one author has experimented with modifications to the standard heuristics: Mészáros [7] describes an inexact version of the minimum local fill ordering.)

One possible remedy for small pivots is diagonal pivoting. At each iteration, a "large" diagonal element is selected from the unreduced portion of the matrix and moved to the pivot position by symmetric row and column pivoting. The algorithm is terminated when none of the remaining diagonal elements is sufficiently large, and an approximate solution is computed with the partial factors. (See Higham [4, Chapter 10] for details and error analysis.) This strategy is not particularly appealing in the context of interior-point linear programming codes because of the loss of efficiency due to shifting of data during the numerical factorization. Moreover, there is little incentive to test this strategy because the simple patches described above perform so well in practice.

In this article, we use standard results from numerical analysis to explain the good performance of these patching strategies on the vast majority of problems. We also gain some insight into their limitations and into how and why they fail.

Our error analysis for the modified Cholesky algorithm is rigorous, with explicitly stated assumptions and precise bounds (see Sections 3 and 4). We revert, however, to a more informal style when applying these results to the interior-point context (Section 5). The reason is pure pragmatism. A fully rigorous analysis would be impossibly complex, notationally speaking, and unduly pessimistic. The informal analysis yields adequate insight into the typical performance of the algorithm, as our computational results in Section 6 demonstrate.

A number of other papers on linear algebra operations in barrier and interior-point methods have appeared in recent years. Wright [12] has considered the Newton-logarithmic barrier method for general constrained optimization, in which the linear system to be solved for the Newton step is positive semidefinite and ill conditioned during later iterations. She uses a Cholesky factorization with diagonal pivoting to identify the subspace spanned by the active constraint Jacobian. From this information, an accurate solution of the Newton equations can be obtained, in which the components of the step in both the range space of the active constraint Jacobian and the null space of its transpose are well resolved. Our analysis has a similar flavor to Wright's, but the application is somewhat different. The unknowns in our linear system are the unconstrained dual variables rather than the primals and, since this problem is linear, we have little interest in resolving the component of the step in the near-null space of the coefficient matrix. We focus too on Cholesky algorithms that perform no pivoting during the numerical factorization, reflecting computational practice in the current generation of interior-point linear programming codes.

In an earlier paper [14], we considered the stability of algorithms for the symmetric indefinite form of the step equations at each iteration of a interior-point method for linear programming. We showed that, despite the ill-conditioning of the coefficient matrix, the steps obtained by this approach are good search directions for the interior-point method. Forsgren, Gill, and Shinnerl [2] perform a similar analysis in the context of logarithmic barrier methods.

The remainder of this paper is organized as follows. In Section 2, we introduce primal-dual interior-point methods and derive the linear equations to be solved at each iteration of these methods. Section 3 introduces Algorithm **modchol**, the modified Cholesky procedure, and examines the accuracy of the solution obtained with this factorization, under certain assumptions on the eigenvalues of the factored matrix. In Section 4, we account for the effect of finite-precision floating-point arithmetic on solution accuracy. We return to the interior-point application in Section 5, showing that Algorithm **modchol** yields good steps for these methods until the duality gap becomes very small, even if the linear program is primal or dual degenerate. The analytical results are verified by computational experiments with an interior-point code using Algorithm **modchol**, which are reported in Section 6.

Notation. We summarize here the notation used in the remainder of the paper.

The ith singular value of a matrix A is denoted by $\sigma_i(A)$. We use σ_i alone to denote the ith singular value of the exact Cholesky factor L in Section 3.

For any matrix M and index steps $\mathcal{I}$ and $\mathcal{J}$, $M_{\mathcal{I}\mathcal{J}}$ denotes the submatrix formed by the elements M_{ij}, for $i \in \mathcal{I}$ and $j \in \mathcal{J}$. The ith column of M is denoted by $M_{\cdot i}$, and the column submatrix consisting of columns $j \in \mathcal{J}$ is denoted by $M_{\cdot \mathcal{J}}$.

Unit roundoff error is denoted by $\mathbf{u}$. Higham [4, Chapter 1] defines $\mathbf{u}$ implicitly by the statement that when α and ζ are any two floating-point numbers, op denotes $+$, $-$, $\times$, and $/$, and $fl(\cdot)$ denotes the floating-point representation of a real number, we have

$$fl(\alpha \,\mathrm{op}\, \zeta) = (\alpha \,\mathrm{op}\, \zeta)(1+\delta) \quad \text{for some } \delta \text{ satisfying } |\delta| \le \mathbf{u}.$$

For any positive integer m with $m\mathbf{u} < 1$, we define

$$\gamma_m = \frac{m\mathbf{u}}{1 - m\mathbf{u}} \tag{1}$$

(see Higham [4, Lemma 2.1]).

The notation $\|\cdot\|$ denotes the Euclidean vector norm $\|\cdot\|_2$ and also its induced matrix norm, unless otherwise noted. For any matrix A, the matrix consisting of the absolute values of each element is denoted by $|A|$. We use $\mathbf{1}$ to denote the vector $(1, 1, \cdots, 1)^T$.

Finally, we mention the parameter ϵ that defines the pivot threshold in the modified Cholesky algorithm. A second quantity $\bar{\epsilon}$, which is related to ϵ by

$$\bar{\epsilon} \overset{\mathrm{def}}{=} 2m^2\epsilon,$$

appears frequently in the analysis because the incorporation of the scaling term $2m^2$ saves notational clutter.

2. Primal-Dual Algorithms for Linear Programming. We consider the linear programming problem in standard form:

$$\min\ c^T x \quad \text{subject to} \quad Ax = b, \qquad x \ge 0, \tag{2}$$

where $x \in \mathsf{R}^n$, $c \in \mathsf{R}^n$, $A \in \mathsf{R}^{m\times n}$, and $b \in \mathsf{R}^m$. The dual of (2) is

$$\max\ b^T \pi \quad \text{subject to} \quad A^T\pi + s = c, \qquad s \ge 0, \tag{3}$$

where $s \in \mathsf{R}^n$ and $\pi \in \mathsf{R}^m$. We assume throughout the paper that A has full row rank, so that $m \le n$. The Karush-Kuhn-Tucker (KKT) conditions, which identify a vector triple (x, π, s) as a primal-dual solution for (2), (3), can be stated as follows:

$$\begin{aligned} &(4a) && A^T\pi + s &=&\ c, \\ &(4b) && Ax &=&\ b, \\ &(4c) && x_i s_i &=&\ 0, \quad i = 1, 2, \ldots, n, \\ &(4d) && (x, s) &\geq&\ 0. \end{aligned}$$

We assume throughout the paper that a primal-dual solution exists. We make no assumptions about uniqueness or nondegeneracy; our analysis in Section 5 continues to hold when the problem (2) is primal or dual degenerate. It is well known that the index set $\{1, 2, \ldots, n\}$ can be partitioned into two sets $\mathcal{B}$ and $\mathcal{N}$ such that for all primal-dual solutions (x^*, π^*, s^*) we have

$$(5) \qquad x_i^* = 0 \quad \text{for all } i \in \mathcal{N}, \qquad s_i^* = 0 \quad \text{for all } i \in \mathcal{B}.$$

Primal-dual interior-point algorithms generate a sequence of iterates (x, π, s) that satisfy the strict inequality $(x, s) > 0$. They find search directions by applying a modification of Newton's method to the system of nonlinear equations formed by the first three KKT conditions (4a),(4b),(4c), namely,

$$(6) \qquad Ax - b = 0, \qquad A^T\pi + s - c = 0, \qquad XS\mathbf{1} = 0,$$

where $X = \text{diag}(x_1, x_2, \ldots, x_n)$, $S = \text{diag}(s_1, s_2, \ldots, s_n)$, and $\mathbf{1} = (1, 1, \ldots, 1)^T$. In general, the search direction $(\Delta x, \Delta\pi, \Delta s)$ is obtained from the following linear system:

$$(7) \qquad \begin{bmatrix} 0 & A^T & I \\ A & 0 & 0 \\ S & 0 & X \end{bmatrix} \begin{bmatrix} \Delta x \\ \Delta\pi \\ \Delta s \end{bmatrix} = \begin{bmatrix} -r_c \\ -r_b \\ -r_{xs} \end{bmatrix},$$

where the coefficient matrix is the Jacobian of (6) and the right-hand side components r_b and r_c are defined by

$$(8) \qquad r_b = Ax - b, \qquad r_c = A^T\pi + s - c.$$

In a pure Newton (affine-scaling) method, the remaining right-hand side component r_{xs} is defined by

$$(9) \qquad r_{xs} = XS\mathbf{1},$$

and, in this case, we denote the solution of (7) by $(\Delta x^{\text{aff}}, \Delta\pi^{\text{aff}}, \Delta s^{\text{aff}})$. In a path-following method, we have

$$(10) \qquad r_{xs} = XS\mathbf{1} - \zeta\mu\mathbf{1},$$

where μ is the duality gap defined by

$$(11) \qquad \mu = x^Ts/n,$$

and $\zeta \in [0, 1]$ is a *centering parameter*. In the "Mehrotra predictor-corrector" algorithm, which is used as the basis of many practical codes, the search direction is calculated by setting

$$(12) \qquad r_{xs} = XS\mathbf{1} + \Delta X^{\text{aff}}\Delta S^{\text{aff}}\mathbf{1} - \zeta\mu\mathbf{1},$$

where ΔX^{aff} and ΔS^{aff} are the diagonal matrices formed from the affine-scaling step components Δx^{aff} and Δs^{aff}. Hence, Mehrotra's method requires the solution of *two* linear systems at each iteration—the affine scaling system (7), (8), (9), and the search direction system (7), (8), (12). A heuristic based on the effectiveness of the affine scaling direction is used to determine the value of ζ in (12).

Once a search direction has been determined, the primal-dual algorithm takes a step of the form

$$(x, \pi, s) + \alpha(\Delta x, \Delta \pi, \Delta s),$$

where α is chosen to maintain strict positivity of the x and s components; that is,

$$(x, s) + \alpha(\Delta x, \Delta s) > 0. \tag{13}$$

In most codes, α is chosen to be some fraction of the step-to-boundary $\alpha_{\max}$ defined as

$$\alpha_{\max} = \sup_{\alpha \in [0,1]} \{\alpha \mid (x, s) + \alpha(\Delta x, \Delta s) \geq 0\}. \tag{14}$$

A typical strategy is to set

$$\alpha = \eta \alpha_{\max},$$

where $\eta \in [.9, 1.0)$ approaches 1 as the interior-point method approaches the solution set.

By applying block elimination to (7) and using the notation

$$D^2 = S^{-1} X, \tag{15}$$

we obtain the following equivalent system:

$$\begin{aligned}
&(16a) & AD^2 A^T \Delta \pi &= -r_b + AD^2 (r_c - X^{-1} r_{xs}), \\
&(16b) & \Delta s &= -r_c - A^T \Delta \pi, \\
&(16c) & \Delta x &= -S^{-1}(r_{xs} + X \Delta s).
\end{aligned}$$

In many codes, the solution is obtained from just this formulation. A sparse Cholesky factorization, modified to handle small pivots, is applied to the symmetric positive definite coefficient matrix AD^2A^T in (16a) and the solution $\Delta \pi$ is obtained by triangular substitution with the computed factor. The remaining direction components are recovered from (16b) and (16c). This technique yields steps $(\Delta x, \Delta \pi, \Delta s)$ that are useful search directions for the interior-point algorithm, *even when the matrix AD^2A^T is ill conditioned*, as often happens during later iterations. This observation is somewhat surprising, since a naive application of error analysis results would suggest that the combination of ill-conditioning and roundoff would corrupt the direction hopelessly. The results of Sections 3, 4, and 5 provide an explanation for this phenomenon.

The following observation is crucial to our analysis: In computing $\Delta \pi$ from (16a), we are not interested so much in the error in $\Delta \pi$ itself as in the effect of this error on the remaining step components Δs and Δx that are recovered from (16b) and (16c), respectively. If the relative errors in these components are large, the positivity requirement (13) may cause the step length α to be significantly shortened, thereby curtailing the algorithm's progress. We return to this issue in Section 5, after describing and analyzing the modified Cholesky algorithm in Sections 3 and 4.

3. A Modified Cholesky Algorithm. In this section, we describe and analyze Algorithm **modchol**, a modified Cholesky algorithm designed to handle ill-conditioned matrices for which small or negative pivots may arise during the factorization.

Algorithm **modchol** accepts an $m \times m$ symmetric positive definite matrix M as input, together with a small positive user-defined parameter ϵ, which defines a threshold of acceptability for the pivot elements. If a candidate pivot element is smaller than this threshold, the algorithm simply skips a step of factorization. Algorithm **modchol** outputs an approximate lower triangular factor $\tilde{L}$ and an index set $\mathcal{J} \subset \{1, 2, \ldots, m\}$ containing the indices of the skipped pivots. In the following specification, we use $M^{(i)}$ to denote the unfactored part of M that remains after i steps of the algorithm.

```
Algorithm modchol

Given $\epsilon$ with $0 < \epsilon \ll 1$;
Set   $M^{(0)} \leftarrow M$; $\tilde{L} \leftarrow 0$; $\mathcal{J} \leftarrow \emptyset$; $\beta = \max_{i=1,2,\ldots,m} M_{ii}$;
for   $i = 1, 2, \ldots, m$
      if    $M_{ii}^{(i-1)} \le \beta\epsilon$
            (* skip this elimination step *)
            Set $\mathcal{J} \leftarrow \mathcal{J} \cup \{i\}$ and
```

$$E^{(i)} = \left[\begin{array}{c|cccc} 0 & 0 & \cdots & \cdots & 0 \\ \hline 0 & M_{ii}^{(i-1)} & \cdots & \cdots & M_{im}^{(i-1)} \\ \vdots & \vdots & 0 & & 0 \\ \vdots & \vdots & & \ddots & \vdots \\ 0 & M_{mi}^{(i-1)} & 0 & \cdots & 0 \end{array}\right], \qquad M^{(i)} = M^{(i-1)} - E^{(i)}; \tag{17}$$

```
      else
            (* perform the usual Cholesky elimination step *)
            $\tilde{L}_{ii} \leftarrow \sqrt{M_{ii}^{(i-1)}}$; $M^{(i)} \leftarrow 0$
            for   $j = i+1, i+2, \ldots, m$
                  $\tilde{L}_{ji} = M_{ij}^{(i-1)} / \tilde{L}_{ii}$ ;
            for   $j = i+1, i+2, \ldots, m$
                  for   $k = i+1, i+2, \ldots, m$
                        $M_{jk}^{(i)} \leftarrow M_{jk}^{(i-1)} - \tilde{L}_{ji}\tilde{L}_{ki}$.
```

The ith column of $\tilde{L}$ is zero for each $i \in \mathcal{J}$; that is, $\tilde{L}_{\cdot\mathcal{J}} = 0$. If we denote

$$E = \sum_{i \in \mathcal{J}} E^{(i)} \tag{18}$$

and denote the complement of $\mathcal{J}$ in $\{1, 2, \ldots, m\}$ by $\bar{\mathcal{J}}$, it follows from (17) that

$$E_{\bar{\mathcal{J}}\bar{\mathcal{J}}} = 0. \tag{19}$$

That is, the row or column index of each nonzero element in E must lie in $\mathcal{J}$. It follows from the algorithm that $\tilde{L}$ is the exact Cholesky factor of the perturbed matrix $M - E$, which we denote for convenience by $\tilde{M}$. That is, we have

$$\tilde{L}\tilde{L}^T = \tilde{M} = M - E. \tag{20}$$

By partitioning this equation into its $\mathcal{J}$ and $\bar{\mathcal{J}}$ components and using $\tilde{L}_{\cdot\mathcal{J}} = 0$ and (19), we obtain

$$\text{(21a)} \qquad M_{\bar{\mathcal{J}}\bar{\mathcal{J}}} = \tilde{L}_{\bar{\mathcal{J}}\cdot}\tilde{L}^T_{\bar{\mathcal{J}}\cdot} + E_{\bar{\mathcal{J}}\bar{\mathcal{J}}} = \tilde{L}_{\bar{\mathcal{J}}\bar{\mathcal{J}}}\tilde{L}^T_{\bar{\mathcal{J}}\bar{\mathcal{J}}}$$

$$\text{(21b)} \qquad M_{\bar{\mathcal{J}}\mathcal{J}} = \tilde{L}_{\bar{\mathcal{J}}\cdot}\tilde{L}^T_{\mathcal{J}\cdot} + E_{\bar{\mathcal{J}}\mathcal{J}} = \tilde{L}_{\bar{\mathcal{J}}\bar{\mathcal{J}}}\tilde{L}^T_{\mathcal{J}\bar{\mathcal{J}}} + E_{\bar{\mathcal{J}}\mathcal{J}}.$$

The *exact* Cholesky factor L (whose existence is guaranteed by the assumed positive definiteness of M) satisfies

$$\text{(22)} \qquad LL^T = M.$$

Given the linear system

$$\text{(23)} \qquad Mz = r,$$

where M is the matrix factored by **modchol**, the exact solution obviously satisfies

$$\text{(24)} \qquad z = M^{-1}r = L^{-T}L^{-1}r.$$

The approximate solution $\tilde{z}$ is chosen so that the partial vector $\tilde{z}_{\bar{\mathcal{J}}}$ solves the reduced system $M_{\bar{\mathcal{J}}\bar{\mathcal{J}}}\tilde{z}_{\bar{\mathcal{J}}} = r_{\bar{\mathcal{J}}}$, while the complementary subvector $z_{\mathcal{J}}$ is set to zero. From (21a), we see that $\tilde{z}_{\bar{\mathcal{J}}}$ can be calculated by performing a pair of triangular substitutions; that is,

$$\text{(25)} \qquad \tilde{z}_{\bar{\mathcal{J}}} = \tilde{L}^{-T}_{\bar{\mathcal{J}}\bar{\mathcal{J}}}\tilde{L}^{-1}_{\bar{\mathcal{J}}\bar{\mathcal{J}}}r_{\bar{\mathcal{J}}}, \qquad \tilde{z}_{\mathcal{J}} = 0.$$

Note that $z = \tilde{z}$ when $\mathcal{J} = \emptyset$. When $\mathcal{J} \neq 0$, on the other hand, the difference between $\tilde{z}$ and z can be large in a relative sense. We have

$$\|z - \tilde{z}\| = \left\| \begin{bmatrix} z_{\mathcal{J}} - 0 \\ z_{\bar{\mathcal{J}}} - \tilde{z}_{\bar{\mathcal{J}}} \end{bmatrix} \right\| \geq \|z_{\mathcal{J}}\|,$$

and there is no reason to expect $z_{\mathcal{J}}$ to be small with respect to the full vector z. We can show, however, that the difference between $L^T z$ and $L^T\tilde{z}$ is relatively small under certain assumptions; this result is the culmination of the analysis of this section (Theorem 3.6). As we see in Section 5, this difference determines the usefulness of the computed solution of (16) as a search direction for the interior-point algorithm.

To simplify the analysis, we assume implicitly throughout the paper that

$$\text{(26)} \qquad \beta = 1.$$

A trivial scaling, which affects neither the algorithm nor its analysis, can always be applied to the symmetric positive definite matrix M to yield (26).

We start with a sequence of three results that lead to a bound on the difference between $\tilde{L}^T z$ and $\tilde{L}^T\tilde{z}$. These results require few assumptions on the matrix M and are relatively simple to prove.

LEMMA 3.1. *The submatrix formed by the last $m - i$ rows and columns of $M^{(i)}$ is symmetric positive definite, for all $i = 0, 1, \ldots, m-1$. Moreover, the diagonal elements of all these submatrices are bounded by* 1.

Proof. This observation follows by a simple inductive argument. By assumption, the starting matrix $M^{(0)} = M$ is positive definite. Suppose that the desired property holds for $M^{(i-1)}$. If $i \in \mathcal{J}$, then the lower right $(m-i) \times (m-i)$ submatrix of

$M^{(i)}$ is identical to the lower right $(m-i) \times (m-i)$ submatrix of $M^{(i-1)}$, which is positive definite by assumption. Otherwise, if $i \notin \mathcal{J}$, then $M^{(i)}$ is obtained by applying one step of Cholesky reduction to $M^{(i-1)}$. It is known that the remaining submatrix resulting from this operation is positive definite; hence, the lower right $(m-i) \times (m-i)$ submatrix in question is positive definite, and the desired property holds.

The second claim follows immediately from the fact that $M_{ii} \le \beta = 1$, $i = 1, 2, \ldots, m$ and the fact that the diagonal elements cannot increase during Algorithm **modchol**. □

LEMMA 3.2. *For each $i \in \mathcal{J}$, we have*

$$||E^{(i)}||_2 \le ||E^{(i)}||_F \le (2m\epsilon)^{1/2}.$$

Therefore,

$$||E||_2 \le ||E||_F \le \bar{\epsilon}^{1/2}, \tag{27}$$

where $\bar{\epsilon} = 2m^2\epsilon$.

Proof. From Lemma 3.1, we have $(M_{i,l}^{(i-1)})^2 \le M_{i,i}^{(i-1)} M_{l,l}^{(i-1)}$ for each $l = i+1, \ldots, m$. Suppose $i \in \mathcal{J}$, so that $M_{i,i}^{(i-1)} \le \epsilon$. Since the diagonals of each submatrix $M^{(i-1)}$ are bounded by 1, we have $M_{l,l}^{(i-1)} \le 1$ and therefore

$$\left| M_{i,l}^{(i-1)} \right| \le \left(M_{i,i}^{(i-1)} M_{l,l}^{(i-1)} \right)^{1/2} \le \epsilon^{1/2}, \qquad l = i+1, \ldots, m.$$

Hence, we have

$$||E^{(i)}||_2^2 \le ||E^{(i)}||_F^2 \le (M_{i,i}^{(i-1)})^2 + 2 \sum_{l=i+1}^{m} (M_{i,l}^{(i-1)})^2 \le \epsilon^2 + 2(m-i)\epsilon \le 2m\epsilon,$$

thereby proving the first claim. By (18), we have

$$||E||_F^2 = \sum_{i \in \mathcal{J}} ||E^{(i)}||_F^2 \le |\mathcal{J}| 2m\epsilon \le 2m^2\epsilon,$$

thereby proving (27). □

In the case in which all the small pivots appear in the bottom right corner of the matrix (that is, $\mathcal{J} = \{p+1, p+2, \ldots, m\}$ for some index p), the estimate (27) can be improved to

$$||E||_2 \le ||E||_F \le m^2\epsilon = .5\bar{\epsilon}, \tag{28}$$

This stronger estimate applies in most instances of the interior-point application of Section 5.

We are now able to derive an estimate of the difference between $\tilde{L}^T z$ and $\tilde{L}^T \tilde{z}$.

THEOREM 3.3. *For the exact solution z and approximate solution $\tilde{z}$ defined in (24) and (25), respectively, we have that*

$$||\tilde{L}^T[z - \tilde{z}]|| = ||\tilde{L}_{\bar{\mathcal{J}}\bar{\mathcal{J}}}^{-1} E_{\bar{\mathcal{J}}\mathcal{J}} z_{\mathcal{J}}|| \le ||\tilde{L}_{\bar{\mathcal{J}}\bar{\mathcal{J}}}^{-1}|| \, ||E_{\bar{\mathcal{J}}\mathcal{J}}|| \, ||z_{\mathcal{J}}||. \tag{29}$$

Proof. From (24) together with (21), we have

$$\begin{aligned} r_{\bar{\mathcal{J}}} &= M_{\bar{\mathcal{J}}\bar{\mathcal{J}}} z_{\bar{\mathcal{J}}} + M_{\bar{\mathcal{J}}\mathcal{J}} z_{\mathcal{J}} \\ &= \tilde{L}_{\bar{\mathcal{J}}\bar{\mathcal{J}}} \tilde{L}^T_{\bar{\mathcal{J}}\bar{\mathcal{J}}} z_{\bar{\mathcal{J}}} + \left[\tilde{L}_{\bar{\mathcal{J}}\bar{\mathcal{J}}} \tilde{L}^T_{\mathcal{J}\bar{\mathcal{J}}} + E_{\bar{\mathcal{J}}\mathcal{J}} \right] z_{\mathcal{J}} \\ &= \tilde{L}_{\bar{\mathcal{J}}\bar{\mathcal{J}}} \tilde{L}^T_{\cdot\bar{\mathcal{J}}} z + E_{\bar{\mathcal{J}}\mathcal{J}} z_{\mathcal{J}}, \end{aligned}$$

while from (25), we have

$$r_{\bar{\mathcal{J}}} = \tilde{L}_{\bar{\mathcal{J}}\bar{\mathcal{J}}} \tilde{L}^T_{\bar{\mathcal{J}}\bar{\mathcal{J}}} \tilde{z}_{\bar{\mathcal{J}}} = \tilde{L}_{\bar{\mathcal{J}}\bar{\mathcal{J}}} \left[\tilde{L}^T_{\bar{\mathcal{J}}\bar{\mathcal{J}}} \tilde{z}_{\bar{\mathcal{J}}} + \tilde{L}_{\mathcal{J}\bar{\mathcal{J}}} \tilde{z}_{\mathcal{J}} \right] = \tilde{L}_{\bar{\mathcal{J}}\bar{\mathcal{J}}} \tilde{L}^T_{\cdot\bar{\mathcal{J}}} \tilde{z}.$$

By combining these two relations, we obtain

$$\tilde{L}^T_{\cdot\bar{\mathcal{J}}} [z - \tilde{z}] = -\tilde{L}^{-1}_{\bar{\mathcal{J}}\bar{\mathcal{J}}} E_{\bar{\mathcal{J}}\mathcal{J}} z_{\mathcal{J}}. \tag{30}$$

Since $\tilde{L}_{\cdot\mathcal{J}} = 0$, the result follows immediately. □

The remaining analysis of this section requires some additional assumptions on the distribution of the singular values of M and on the parameter ϵ. Accordingly, we introduce a little more notation. The eigenvalues of M are denoted by σ_i^2, $i = 1, 2, \ldots, m$, where

$$\sigma_1^2 \geq \sigma_2^2 \geq \cdots \geq \sigma_m^2 > 0. \tag{31}$$

We define the diagonal matrix Σ by

$$\Sigma = \operatorname{diag}(\sigma_1, \sigma_2, \ldots, \sigma_m). \tag{32}$$

It follows that there exists an orthogonal matrix Q such that

$$M = Q \Sigma^2 Q^T. \tag{33}$$

Because the largest diagonal in M is 1, we have by elementary analysis that

$$1 \leq \sigma_1^2 \leq m. \tag{34}$$

In the subsequent analysis, we assume that there is an integer p with $1 \leq p \leq m$ such that

- ϵ is small relative to σ_p^2; and
- if $p < m$, there is a significant gap in the spectrum of M between σ_p^2 and σ_{p+1}^2.

(We will be more specific about these two assumptions presently.) By partitioning the spectrum at the gap, we obtain

$$\Sigma_1 = \operatorname{diag}(\sigma_1, \sigma_2, \ldots, \sigma_p), \qquad \Sigma_2 = \operatorname{diag}(\sigma_{p+1}, \sigma_{p+2}, \ldots, \sigma_m). \tag{35}$$

From (33), Q can be partitioned accordingly to obtain

$$Q = [Q_1 \,|\, Q_2], \qquad M = Q_1 \Sigma_1^2 Q_1^T + Q_2 \Sigma_2^2 Q_2^T.$$

Since $M = LL^T$, it follows that σ_i, $i = 1, 2, \ldots, m$ are the singular values of L. In fact, we must have

$$L^T = U \Sigma Q^T \tag{36}$$

for some $m \times m$ orthogonal matrix U, where Σ and Q are defined as above.

We use $\tilde{\sigma}_i^2$, $i = 1, 2, \ldots, m$ to denote the eigenvalues of the perturbed matrix $\tilde{M}$. It follows immediately from (20) that the singular values of $\tilde{L}$ are $\tilde{\sigma}_i$, $i = 1, 2, \ldots, m$. The rank of $\tilde{L}$ is $|\bar{\mathcal{J}}|$, because $\tilde{L}_{\bar{\mathcal{J}}\bar{\mathcal{J}}}$ is lower triangular with nonzero diagonals while $\tilde{L}_{\cdot\mathcal{J}} = 0$. Therefore, we have

$$\tilde{\sigma}_{|\bar{\mathcal{J}}|} > \tilde{\sigma}_{|\bar{\mathcal{J}}|+1} = \cdots = \tilde{\sigma}_m = 0. \tag{37}$$

As in (36), there are orthogonal $m \times m$ matrices $\tilde{U}$ and $\tilde{Q}$ such that

$$\tilde{L}^T = \tilde{U}\tilde{\Sigma}\tilde{Q}^T, \tag{38}$$

where

$$\tilde{\Sigma} = \mathrm{diag}(\tilde{\sigma}_1, \tilde{\sigma}_2, \ldots, \tilde{\sigma}_m).$$

It is an immediate consequence of an eigenvalue perturbation result of Stewart and Sun [10, Corollary IV.4.13] and Lemma 3.2 that

$$\left\{ \sum_{i=1}^{m} [\sigma_i^2 - \tilde{\sigma}_i^2]^2 \right\}^{1/2} \leq \|E\|_F = \bar{\epsilon}^{1/2}. \tag{39}$$

The main assumption of this section is that $|\bar{\mathcal{J}}| = p$; that is, Algorithm **modchol** correctly identifies the numerical rank of the matrix M. One might expect that we should not have to *assume* this equality at all—that it should follow from the spectrum gap and from a judicious choice of ϵ. Practical experience supports this expectation; the algorithm has little trouble determining the numerical rank on the vast majority of problems. In fact, part of the result—the bound $|\bar{\mathcal{J}}| \geq p$—follows from a minimal assumption on ϵ.

Lemma 3.4. *If $\bar{\epsilon}^{1/2} < \sigma_p^2$, we have $|\bar{\mathcal{J}}| \geq p$.*

Proof. If $|\bar{\mathcal{J}}| < p$, we have from (37) and (39) that

$$\sigma_p^2 \leq \sigma_{|\bar{\mathcal{J}}|+1}^2 = \left| \sigma_{|\bar{\mathcal{J}}|+1}^2 - \tilde{\sigma}_{|\bar{\mathcal{J}}|+1}^2 \right| \leq \bar{\epsilon}^{1/2},$$

contradicting our assumption that $\bar{\epsilon}^{1/2} < \sigma_p^2$. □

However, the conditions on ϵ, σ_p, and σ_{p+1} needed to prove the other half of the result—$|\bar{\mathcal{J}}| \leq p$—are too rigorous to be useful. This is a consequence of the fact that poorly conditioned triangular matrices need not have particularly small diagonal elements (see Lawson and Hanson [5, p. 31] for the classic example of this phenomenon).

Our next result concerns perturbation of the subspace spanned by Q_1, which is the invariant subspace of "large" eigenvalues of M.

Lemma 3.5. *Suppose that $|\bar{\mathcal{J}}| = p < m$ and that the values σ_p and σ_{p+1} from (31) and ϵ from Lemma 3.2 satisfy the conditions*

$$\frac{\sigma_{p+1}^2}{\sigma_p^2} \leq .1, \tag{40a}$$

$$\sigma_p^2 - \sigma_{p+1}^2 > 5\bar{\epsilon}^{1/2}. \tag{40b}$$

Then there is a $p \times p$ symmetric positive definite matrix $\tilde{\Lambda}$ and an orthonormal $m \times p$ matrix $\tilde{Q}_1$ such that

$$\tilde{M} = \tilde{Q}_1 \tilde{\Lambda} \tilde{Q}_1^T, \tag{41}$$

$$\|\tilde{Q}_1 - Q_1\| \leq \frac{2\bar{\epsilon}^{1/2}}{\sigma_p^2 - \sigma_{p+1}^2 - 2\bar{\epsilon}^{1/2}}, \tag{42}$$

$$\|\tilde{\Lambda} - \Sigma_1^2\| \leq 2\bar{\epsilon}^{1/2}. \tag{43}$$

(The constants used in (40a) and in similar expressions should not be taken too seriously. We assign them specific values only to avoid an excess of notation.)

Proof. The result is a straightforward consequence of Theorem V.2.8 of Stewart and Sun [10, p. 238]. Since $\tilde{M} = M - E$, we use (33) and partition as in (35) to obtain

$$Q^T \tilde{M} Q = Q^T M Q - Q^T E Q = \begin{bmatrix} \Sigma_1^2 & 0 \\ 0 & \Sigma_2^2 \end{bmatrix} - \begin{bmatrix} F_{11} & F_{12} \\ F_{12}^T & F_{22} \end{bmatrix}.$$

We now make the following identifications with the quantities in the cited result:

$$\tilde{\gamma} = \|F_{12}^T\| \leq \|F\| = \|E\| \leq \bar{\epsilon}^{1/2}, \qquad \tilde{\eta} = \|F_{12}\| \leq \bar{\epsilon}^{1/2},$$
$$\tilde{\delta} = \text{sep}(\Sigma_1^2, \Sigma_2^2) - \|F_{11}\| - \|F_{22}\| \geq \sigma_p^2 - \sigma_{p+1}^2 - 2\bar{\epsilon}^{1/2} > 2\bar{\epsilon}^{1/2},$$

where $\text{sep}(\cdot,\cdot)$ is the minimum distance between the spectra of its two arguments. From the given result, there is a matrix P of dimension $(m-p) \times p$ such that the matrix $\tilde{Q}_1$ defined by

$$\tilde{Q}_1 = Q_1 + Q_2 P \tag{44}$$

is an invariant subspace for $\tilde{M}$, where

$$\|P\| \leq \frac{\tilde{\gamma}}{\tilde{\delta}} \leq \frac{2\bar{\epsilon}^{1/2}}{\sigma_p^2 - \sigma_{p+1}^2 - 2\bar{\epsilon}^{1/2}} < 1. \tag{45}$$

Moreover, the representation of $\tilde{M}$ with respect to $\tilde{Q}_1$ is

$$\tilde{Q}_1^T \tilde{M} \tilde{Q}_1 = \tilde{\Lambda} = \Sigma_1^2 + F_{11} + F_{12} P. \tag{46}$$

The bound (42) follows from (44), (45), and $\|Q_2\| = 1$. It follows immediately from the first equality in (46) that $\tilde{\Lambda}$ is symmetric, and we have

$$\|\tilde{\Lambda} - \Sigma_1^2\| \leq \|F_{11}\| + \|F_{12}\| \|P\| \leq 2\bar{\epsilon}^{1/2},$$

verifying the inequality (43). This inequality implies that the smallest singular value of $\tilde{\Lambda}$ is no smaller than $\sigma_p^2 - 2\bar{\epsilon}^{1/2} > 0$, so $\tilde{\Lambda}$ is symmetric positive definite.

The cited result states further that the matrix $\tilde{Q}_2 = Q_2 - Q_1 P^T$ is orthogonal to $\tilde{Q}_1$ and also defines an invariant subspace for $\tilde{M}$. In fact, we have

$$[\tilde{Q}_1 \,|\, \tilde{Q}_2]^T \tilde{M} [\tilde{Q}_1 \,|\, \tilde{Q}_2] = \begin{bmatrix} \tilde{\Lambda} & 0 \\ 0 & \hat{\Lambda} \end{bmatrix},$$

for some $(m-p) \times (m-p)$ symmetric matrix $\hat{\Lambda}$. Since $\tilde{\Lambda}$ and $\tilde{M}$ both have rank b, we must have $\hat{\Lambda} = 0$, so we have

$$\tilde{M} = [\tilde{Q}_1 \,|\, \tilde{Q}_2] \begin{bmatrix} \tilde{\Lambda} & 0 \\ 0 & 0 \end{bmatrix} [\tilde{Q}_1 \,|\, \tilde{Q}_2]^T = \tilde{Q}_1 \tilde{\Lambda} \tilde{Q}_1^T.$$

Hence, (41) is also satisfied, and the proof is complete. □

Combining (40b) with (39), we obtain

$$\tilde{\sigma}_1^2 \le \sigma_1^2 + \bar{\epsilon}^{1/2} < \sigma_1^2 + .2\sigma_p^2 < 2\sigma_1^2. \tag{47}$$

Another quantity that enters into our error bounds is the norm of $\tilde{L}^{-1}_{\bar{\mathcal{J}}\bar{\mathcal{J}}}$. We denote

$$\tau \overset{\text{def}}{=} \max(\|\tilde{L}^{-1}_{\bar{\mathcal{J}}\bar{\mathcal{J}}}\|, 1) = \max\left(\sigma_{|\bar{\mathcal{J}}|}(\tilde{L}_{\bar{\mathcal{J}}\bar{\mathcal{J}}})^{-1}, 1\right), \tag{48}$$

where $\sigma_{|\bar{\mathcal{J}}|}(\tilde{L}_{\bar{\mathcal{J}}\bar{\mathcal{J}}})$ denotes the $|\bar{\mathcal{J}}|$th singular value of $\tilde{L}_{\bar{\mathcal{J}}\bar{\mathcal{J}}}$. (The lower bound of 1 in (48) simplifies our analysis.) Note from (21a) that

$$\|M^{-1}_{\bar{\mathcal{J}}\bar{\mathcal{J}}}\| = \|\tilde{L}^{-1}_{\bar{\mathcal{J}}\bar{\mathcal{J}}}\|^2 \le \tau^2. \tag{49}$$

Since $\|M_{\bar{\mathcal{J}}\bar{\mathcal{J}}}\| \le \|M\| \le \sigma_1^2$, we have from (34) and (49) that

$$\kappa(M_{\bar{\mathcal{J}}\bar{\mathcal{J}}}) \le \sigma_1^2\tau^2 \le m\tau^2. \tag{50}$$

Under the assumption $|\bar{\mathcal{J}}| = p$, the nonzero part of $\tilde{L}$—the submatrix $\tilde{L}_{\cdot\bar{\mathcal{J}}}$—has full rank p and singular values $\tilde{\sigma}_1, \ldots, \tilde{\sigma}_p$. Since $\tilde{L}_{\cdot\bar{\mathcal{J}}}$ differs from $\tilde{L}_{\bar{\mathcal{J}}\bar{\mathcal{J}}}$ in the presence of the additional rows $\tilde{L}_{\mathcal{J}\bar{\mathcal{J}}}$, we have

$$\sigma_{|\bar{\mathcal{J}}|}(\tilde{L}_{\bar{\mathcal{J}}\bar{\mathcal{J}}}) = \sigma_p(\tilde{L}_{\bar{\mathcal{J}}\bar{\mathcal{J}}}) \le \tilde{\sigma}_p,$$

and therefore

$$\tau\tilde{\sigma}_p \ge 1.$$

The additional rows $\tilde{L}_{\mathcal{J}\bar{\mathcal{J}}}$ can have nontrivial magnitude relative to $\tilde{L}_{\bar{\mathcal{J}}\bar{\mathcal{J}}}$, so $\tilde{\sigma}_p$ may be significantly larger than τ^{-1}. However, $\tilde{\sigma}_p$ cannot be *too* large, since from (39), (40b), and (34), we have that

$$\tilde{\sigma}_p^2 \le \sigma_p^2 + \bar{\epsilon}^{1/2} \le 1.2\sigma_p^2 \le 1.2\sigma_1^2 \le 1.2m.$$

For the purposes of our analysis, we make the assumption that $\tau\tilde{\sigma}_p^2$ is moderate in size. Specifically, we assume that

$$\tau\tilde{\sigma}_p^2 \le 10. \tag{51}$$

Because of (39) and (40b), we have $\tilde{\sigma}_p^2 \ge \sigma_p^2 - \bar{\epsilon}^{1/2} \ge .8\sigma_p^2$, so (51) implies that

$$\tau\sigma_p^2 \le 1.25\tau\tilde{\sigma}_p^2 \le 13.$$

and, in addition,

$$\tau\bar{\epsilon}^{1/2} \le .2\tau\sigma_p^2 \le .25\tau\tilde{\sigma}_p^2 \le 3. \tag{52}$$

We can now prove the main result of this section.

THEOREM 3.6. *Suppose that $|\bar{\mathcal{J}}| = p < m$, that the conditions (40) hold, and that the estimate (51) is satisfied. We then have*

$$\|L^T(\tilde{z} - z)\| \le \left(56\frac{\sigma_1^3}{\sigma_p^3}\bar{\epsilon}^{1/2} + 6\sigma_1\sigma_{p+1}\right)\tau\|z_{\mathcal{J}}\|.$$

Proof. From (36), we have

$$\|L^T(\tilde{z}-z)\| = \|U\Sigma Q^T(\tilde{z}-z)\| = \|\Sigma Q^T(\tilde{z}-z)\|,$$

since U is orthogonal. Now from the partition (35), and using the fact that $\|Q_2\| = 1$ (unless of course Σ_2 and Q_2 are vacuous), we obtain

$$\begin{aligned}
&\|L^T(\tilde{z}-z)\| \\
&\quad \leq \|\Sigma_1 Q_1^T(\tilde{z}-z)\| + \|\Sigma_2\|\,\|\tilde{z}-z\| \\
&\quad \leq \|\Sigma_1^{-1}\|\,\|\Sigma_1^2 Q_1^T(\tilde{z}-z)\| + \|\Sigma_2\|\,\|\tilde{z}-z\| \\
&\quad \leq \|\Sigma_1^{-1}\|\,\|\tilde{\Lambda}\tilde{Q}_1^T(\tilde{z}-z)\| + \|\Sigma_1^{-1}\|\,\|\tilde{\Lambda}\tilde{Q}_1^T - \Sigma_1^2 Q_1^T\|\,\|\tilde{z}-z\| + \|\Sigma_2\|\,\|\tilde{z}-z\|. \qquad (53)
\end{aligned}$$

The first term in this expression is easiest to bound. From (35), we have $\|\Sigma_1^{-1}\| = \sigma_p^{-1}$. Applying the relations (41), (20), (38), (47), (29), (27), and (48), respectively, we obtain

$$\begin{aligned}
\|\tilde{\Lambda}\tilde{Q}_1^T(\tilde{z}-z)\| &= \|\tilde{M}(\tilde{z}-z)\| \\
&= \|\tilde{L}\tilde{L}^T(\tilde{z}-z)\| \\
&\leq \tilde{\sigma}_1\|\tilde{L}^T(\tilde{z}-z)\| \\
&\leq 2\sigma_1\|\tilde{L}^T(\tilde{z}-z)\| \\
&\leq 2\sigma_1\|\tilde{L}_{\bar{\mathcal{J}}\bar{\mathcal{J}}}^{-1}\|\,\|E_{\bar{\mathcal{J}}\mathcal{J}}\|\,\|z_{\mathcal{J}}\| \\
&\leq 2\sigma_1\tau\bar{\epsilon}^{1/2}\|z_{\mathcal{J}}\|.
\end{aligned}$$

We therefore have

$$\|\Sigma_1^{-1}\|\,\|\tilde{\Lambda}\tilde{Q}_1^T(\tilde{z}-z)\| \leq 2\frac{\sigma_1}{\sigma_p}\tau\bar{\epsilon}^{1/2}\|z_{\mathcal{J}}\|. \tag{54}$$

The second and third terms in (53) require a bound on $\|\tilde{z}-z\|$. From (30) and the fact that $\tilde{z}_{\mathcal{J}} = 0$, we have

$$\tilde{L}_{\bar{\mathcal{J}}\bar{\mathcal{J}}}^T(\tilde{z}_{\bar{\mathcal{J}}} - z_{\bar{\mathcal{J}}}) = \tilde{L}_{\mathcal{J}\bar{\mathcal{J}}}^T z_{\mathcal{J}} + \tilde{L}_{\bar{\mathcal{J}}\bar{\mathcal{J}}}^{-1} E_{\bar{\mathcal{J}}\mathcal{J}} z_{\mathcal{J}}, \tag{55}$$

and therefore

$$\|\tilde{z}_{\bar{\mathcal{J}}} - z_{\bar{\mathcal{J}}}\| \leq \|\tilde{L}_{\bar{\mathcal{J}}\bar{\mathcal{J}}}^{-T}\|\left(\|\tilde{L}_{\mathcal{J}\bar{\mathcal{J}}}\| + \|\tilde{L}_{\bar{\mathcal{J}}\bar{\mathcal{J}}}^{-1}\|\,\|E_{\bar{\mathcal{J}}\mathcal{J}}\|\right)\|z_{\mathcal{J}}\|. \tag{56}$$

From (47), we have $\|\tilde{L}_{\mathcal{J}\bar{\mathcal{J}}}\| \leq \tilde{\sigma}_1 \leq 2\sigma_1$, while from (27), we have $\|E_{\bar{\mathcal{J}}\mathcal{J}}\| \leq \bar{\epsilon}^{1/2}$. Substituting these estimates into (56) and using (52), we obtain

$$\|\tilde{z}_{\bar{\mathcal{J}}} - z_{\bar{\mathcal{J}}}\| \leq \tau(2\sigma_1 + \tau\bar{\epsilon}^{1/2})\|z_{\mathcal{J}}\| \leq (2\sigma_1\tau + 3\tau)\|z_{\mathcal{J}}\|. \tag{57}$$

Finally, using $\tilde{z}_{\mathcal{J}} = 0$ together with $\tau \geq 1$, (34), and (57), we obtain

$$\|\tilde{z}-z\| \leq \|\tilde{z}_{\bar{\mathcal{J}}} - z_{\bar{\mathcal{J}}}\| + \|z_{\mathcal{J}}\| \leq (2\sigma_1\tau + 3\tau + 1)\|z_{\mathcal{J}}\| \leq 6\sigma_1\tau\|z_{\mathcal{J}}\|. \tag{58}$$

Turning specifically to the second term in (53), we have from (34), Lemma 3.5, (47), and (40) that

$$\|\tilde{\Lambda}\tilde{Q}_1^T - \Sigma_1^2 Q_1^T\| \leq \|\Sigma_1^2 - \tilde{\Lambda}\|\,\|Q_1\| + \|\tilde{\Lambda}\|\,\|Q_1 - \tilde{Q}_1\|$$

$$
\begin{aligned}
&\leq 2\bar{\epsilon}^{1/2} + \tilde{\sigma}_1^2 \frac{2\bar{\epsilon}^{1/2}}{\sigma_p^2 - \sigma_{p+1}^2 - 2\bar{\epsilon}^{1/2}} \\
&\leq 2\bar{\epsilon}^{1/2} \left[1 + \frac{1.2\sigma_1^2}{\sigma_p^2 - \sigma_{p+1}^2 - 2\bar{\epsilon}^{1/2}} \right] \\
&\leq 2\bar{\epsilon}^{1/2} \frac{2.2\sigma_1^2}{\sigma_p^2 - \sigma_{p+1}^2 - 2\bar{\epsilon}^{1/2}} \\
&\leq 4.5\bar{\epsilon}^{1/2} \frac{\sigma_1^2}{\sigma_p^2} \left[1 - \frac{\sigma_{p+1}^2}{\sigma_p^2} - 2\frac{\bar{\epsilon}^{1/2}}{\sigma_p^2} \right]^{-1} \\
&\leq 4.5\bar{\epsilon}^{1/2} \frac{\sigma_1^2}{\sigma_p^2} [1 - .1 - .4]^{-1} \\
&= 9\bar{\epsilon}^{1/2} \frac{\sigma_1^2}{\sigma_p^2}.
\end{aligned}
$$

By combining this bound with (58) and $\|\Sigma_1^{-1}\| = \sigma_p^{-1}$, we obtain

$$
\begin{aligned}
\|\Sigma_1^{-1}\| \, \|\tilde{\Lambda}\tilde{Q}_1^T - \Sigma_1^2 Q_1^T\| \, \|\tilde{z} - z\| &\leq \frac{1}{\sigma_p} \left(9\bar{\epsilon}^{1/2} \frac{\sigma_1^2}{\sigma_p^2} \right) (6\sigma_1 \tau \|z_{\mathcal{J}}\|) \\
&= 54 \frac{\sigma_1^3}{\sigma_p^3} \tau \bar{\epsilon}^{1/2} \|z_{\mathcal{J}}\|. \qquad (59)
\end{aligned}
$$

For the third term in (53), we have from $\|\Sigma_2\| = \sigma_{p+1}$ that

$$
\|\Sigma_2\| \, \|\tilde{z} - z\| \leq 6\sigma_1 \sigma_{p+1} \tau \|z_{\mathcal{J}}\|. \qquad (60)
$$

The result of the theorem is obtained by substituting (54), (59), and (60) into (53). □

Note that if $\mathcal{J} = \emptyset$ (that is, $|\bar{\mathcal{J}}| = m$), we have $\tilde{z} = z$, so the conclusion of Theorem 3.6 holds trivially in this case as well is we define $\sigma_{m+1} = 0$.

4. The Effect of Finite Precision Computations. In the analysis of the preceding section, we assumed for simplicity that all arithmetic was exact. In this section, we take account of the roundoff errors that are introduced when the approximate solution $\tilde{z}$ is calculated in a finite-precision environment.

Our analysis above focused on the approximate solution $\tilde{z}$ obtained from (25), where the subvector $\tilde{z}_{\mathcal{J}}$ satisfies the following system:

$$
M_{\bar{\mathcal{J}}\bar{\mathcal{J}}} \tilde{z}_{\bar{\mathcal{J}}} = \tilde{L}_{\bar{\mathcal{J}}\bar{\mathcal{J}}} \tilde{L}_{\bar{\mathcal{J}}\bar{\mathcal{J}}}^T \tilde{z}_{\bar{\mathcal{J}}} = r_{\bar{\mathcal{J}}}, \qquad (61)
$$

while the subvector $\tilde{z}_{\mathcal{J}}$ is fixed at zero. In this section, we use $\hat{z}$ to denote the finite precision analog of $\tilde{z}$. We examine errors in $\hat{z}$ due to

- roundoff error in Algorithm **modchol**,
- error arising during the triangular substitutions in (61), and
- evaluation error in the right-hand side r.

As we see in Section 5, evaluation error in the right-hand side is a significant feature of the application to interior-point codes. We denote this error by e, so that the right-hand side $r_{\bar{\mathcal{J}}}$ in the system (61) is replaced by $r_{\bar{\mathcal{J}}} + e_{\bar{\mathcal{J}}}$.

Fortunately, our results follow in a straightforward way from existing results for the Cholesky factorization, since a close inspection of Algorithm **modchol** shows that it simply performs a standard Cholesky factorization on the submatrix $M_{\bar{\mathcal{J}}\bar{\mathcal{J}}}$.

Before stating the main results, we introduce two more assumptions. The first concerns the relative sizes of τ and $\mathbf{u}$, specifically,

$$\tau^2\gamma_{m+1} \leq \frac{1}{5m^2}, \tag{62}$$

where γ_{m+1} is defined as in Section 1. Since $\tau > 1$ and $m \geq 1$, it follows immediately that

$$m\gamma_{m+1} \leq .2. \tag{63}$$

The second assumption is that finite precision does not affect cutoff decisions in Algorithm **modchol**. That is, the presence of roundoff error in each submatrix $M^{(i-1)}$ does not affect whether the threshold criterion $M_{ii}^{(i-1)} \leq \beta\epsilon$ passes or fails for each i. This assumption concerns the relative sizes of $\mathbf{u}$ and ϵ, and it requires some explanation. We cannot expect to take care of the "borderline cases" in which some candidate pivots fall just to one side or the other of the threshold. Rather, we want the cases in which there is a clear distinction between small and large pivots in exact arithmetic to retain this distinction in finite precision arithmetic, and we want the threshold $\beta\epsilon$ to fall comfortably inside the "gap" in both settings. In finite precision, the size of rounding error introduced into $M_{ii}^{(i-1)}$ by earlier steps of Algorithm **modchol** is comparable to $\beta\mathbf{u}$. (Each time M_{ii} is updated by the algorithm, a positive number no larger than itself is subtracted from it. Since $|M_{ii}| \leq \beta$, the floating-point error introduced here is bounded by $\beta\mathbf{u}$.) We want these errors to be smaller than the threshold $\beta\epsilon$, so that pivots that are tiny in exact arithmetic do not exceed the threshold in finite precision. Hence, we can state this assumption roughly as follows:

$$\epsilon \geq \mathbf{u}. \tag{64}$$

The following lemma accounts for the effects of finite precision on the approximate solution $\tilde{z}$ obtained from Algorithm **modchol** and (25).

LEMMA 4.1. *Suppose that Algorithm* **modchol** *and the triangular substitutions in (61) are performed in finite-precision arithmetic with perturbed right-hand side $r_{\bar{\mathcal{J}}} + e_{\bar{\mathcal{J}}}$ to yield an approximate solution $\hat{z}$. Suppose, too, that (62) holds and that roundoff error does not affect the composition of $\mathcal{J}$. We then have*

$$\|\tilde{z} - \hat{z}\| \leq 30m^{5/2}\gamma_{m+1}\tau^3\|z\| + 2\tau^2\|e_{\bar{\mathcal{J}}}\|, \tag{65}$$

where z is the exact solution from (23).

Proof. Algorithm **modchol** operates as a standard Cholesky factorization on the submatrix $M_{\bar{\mathcal{J}}\bar{\mathcal{J}}}$, so we can apply a standard perturbation theorem to bound the error in the subvector $\hat{z}_{\bar{\mathcal{J}}}$. From Higham [4, Theorem 10.4], we find that $\hat{z}_{\bar{\mathcal{J}}}$ satisfies

$$(M_{\bar{\mathcal{J}}\bar{\mathcal{J}}} + E^{\mathbf{u}}_{\bar{\mathcal{J}}\bar{\mathcal{J}}})\hat{z}_{\bar{\mathcal{J}}} = r_{\bar{\mathcal{J}}} + e_{\bar{\mathcal{J}}}, \tag{66}$$

where

$$\|E^{\mathbf{u}}_{\bar{\mathcal{J}}\bar{\mathcal{J}}}\| \leq \frac{2m\gamma_{m+1}}{1 - m\gamma_{m+1}}\|M_{\bar{\mathcal{J}}\bar{\mathcal{J}}}\|. \tag{67}$$

Comparing (66) with (61), we find that

$$M_{\bar{\mathcal{J}}\bar{\mathcal{J}}}(\tilde{z}_{\bar{\mathcal{J}}} - \hat{z}_{\bar{\mathcal{J}}}) = E^{\mathbf{u}}_{\bar{\mathcal{J}}\bar{\mathcal{J}}}\hat{z}_{\bar{\mathcal{J}}} - e_{\bar{\mathcal{J}}}.$$

Manipulating in the usual way, we obtain

$$\|\tilde z_{\bar{\mathcal J}} - \hat z_{\bar{\mathcal J}}\| \le \frac{1}{1-\|M^{-1}_{\bar{\mathcal J}\bar{\mathcal J}}\|\,\|E^{\mathbf u}_{\bar{\mathcal J}\bar{\mathcal J}}\|}\|M^{-1}_{\bar{\mathcal J}\bar{\mathcal J}}\|\left(\|E^{\mathbf u}_{\bar{\mathcal J}\bar{\mathcal J}}\|\,\|\tilde z_{\bar{\mathcal J}}\| + \|e_{\bar{\mathcal J}}\|\right). \tag{68}$$

It follows immediately from (67) that

$$\|M^{-1}_{\bar{\mathcal J}\bar{\mathcal J}}\|\,\|E^{\mathbf u}_{\bar{\mathcal J}\bar{\mathcal J}}\| \le \kappa(M_{\bar{\mathcal J}\bar{\mathcal J}})\frac{2m\gamma_{m+1}}{1-m\gamma_{m+1}}, \tag{69}$$

Combining (50), (62), and (63), we obtain

$$\kappa(M_{\bar{\mathcal J}\bar{\mathcal J}})\frac{2m\gamma_{m+1}}{1-m\gamma_{m+1}} \le \frac{m\tau^2(2m\gamma_{m+1})}{.8} \le .5,$$

so that the denominator in (68) is bounded below by .5. Hence, by substitution into (68), using (34), (49), (69), and (63), we have that

$$\begin{aligned}\|\tilde z_{\bar{\mathcal J}} - \hat z_{\bar{\mathcal J}}\| &\le 2\|M^{-1}_{\bar{\mathcal J}\bar{\mathcal J}}\|\left(\|E^{\mathbf u}_{\bar{\mathcal J}\bar{\mathcal J}}\|\,\|\tilde z_{\bar{\mathcal J}}\| + \|e_{\bar{\mathcal J}}\|\right)\\ &\le 2\tau^2\left(m\frac{2m\gamma_{m+1}}{1-m\gamma_{m+1}}\|\tilde z_{\bar{\mathcal J}}\| + \|e_{\bar{\mathcal J}}\|\right)\\ &\le 5m^2\gamma_{m+1}\tau^2\|\tilde z_{\bar{\mathcal J}}\| + 2\tau^2\|e_{\bar{\mathcal J}}\|. \end{aligned}\tag{70}$$

Finally, we bound $\|\tilde z_{\bar{\mathcal J}}\|$ in terms of $\|z\|$. From (34) and (57), we have

$$\|\tilde z_{\bar{\mathcal J}}\| \le \|z_{\bar{\mathcal J}}\| + \|\tilde z_{\bar{\mathcal J}} - z_{\bar{\mathcal J}}\| \le \|z_{\bar{\mathcal J}}\| + (2\sigma_1\tau + 3\tau)\|z_{\mathcal J}\| \le 6\sigma_1\tau\|z\| \le 6m^{1/2}\tau\|z\|.$$

By combining this bound with (70), we obtain the result. □

The major results of Sections 3 and 4 can be summarized in the following theorem.

THEOREM 4.2. *Suppose that Algorithm* **modchol** *and the triangular substitutions in (61) are performed in finite-precision arithmetic with perturbed right-hand side* $r_{\bar{\mathcal J}} + e_{\bar{\mathcal J}}$ *to yield an approximate solution* $\hat z$. *Suppose, too, that (62) holds and that roundoff errors do not affect the composition of* $\mathcal J$. *Finally, suppose that either*

- $\mathcal J = \emptyset$*; or*
- $|\bar{\mathcal J}| = p < m$*, the conditions (40) hold, and the estimate (51) is satisfied.*

We then have

$$\begin{aligned}&\|L^T(\hat z - z)\|\\ &\quad\le \left\{56\frac{\sigma_1^3}{\sigma_p^3}\bar\epsilon^{1/2} + 6\sigma_1\sigma_{p+1} + 30m^3\gamma_{m+1}\tau^2\right\}\tau\|z\| + 2m^{1/2}\tau^2\|e_{\bar{\mathcal J}}\|.\end{aligned}\tag{71}$$

Proof. When $\mathcal J = \emptyset$, the result is immediate from Lemma 4.1 and $\tilde z = z$. For the remaining case, we obtain (71) by combining the results of Theorem 3.6 and Lemma 4.1. We need note only that $\|z_{\mathcal J}\| \le \|z\|$ and that, from (34), we have

$$\|L^T(\tilde z - \hat z)\| \le \|L\|\,\|\tilde z - \hat z\| = \sigma_1\|\tilde z - \hat z\| \le m^{1/2}\|\tilde z - \hat z\|.$$

□

5. Application to the Interior-Point Algorithm.

In this section, we return to the motivating application: primal-dual interior-point software for linear programming and, in particular, the linear system (16) that is solved at each iteration. We apply the main result—Theorem 4.2—and examine the effect of the parameter ϵ and unit roundoff $\mathbf{u}$ on the quality of the computed search direction $(\widehat{\Delta x}, \widehat{\Delta \pi}, \widehat{\Delta s})$. Our focus is on the later iterations of the interior-point algorithm, during which μ is small and the ill-conditioning of AD^2A^T can become acute. Our results show how and why errors arise in $(\widehat{\Delta x}, \widehat{\Delta \pi}, \widehat{\Delta s})$ and what effect these errors have on the step length, the convergence of the algorithm, and the accuracy that can be attained by this algorithm. They also suggest an appropriate size for the parameter ϵ.

In this section, we revert to an informal style of analysis, using order notation to hide constants of moderate size. Thus if η and ζ are two positive numbers, we write $\eta = O(\zeta)$ if the ratio η/ζ is not too large. Similarly, we write $\eta = \Omega(\zeta)$ if $\eta = O(\zeta)$ and $\zeta = O(\eta)$. Conventionally, order notation is used only when η and ζ are quantities that approach zero in the limit of the algorithm in question. Here, however, we use it in connection with the unit roundoff $\mathbf{u}$, which is small *but fixed.* This slight abuse of notation results in a much clearer insight into the behavior of Algorithm **modchol** in the interior-point context.

In the next subsection, we look closely at the affine-scaling step, for which r_{xs} is defined by (9). This step is important because it closely approximates the steps taken by most rapidly converging algorithms during their final iterations. Subsection 5.2 shows that the steps calculated during the final stages of Mehrotra's predictor corrector algorithm (and therefore by most interior-point codes) have essentially the same properties as affine-scaling steps.

5.1. Affine-Scaling Steps.

We start by estimating the sizes of the various constituents of the equations (16)—the residuals r_b and r_c, the $\mathcal{B}$ and $\mathcal{N}$ components of x, s, and the diagonal matrix D. In standard infeasible-interior-point algorithms (see, for example, Wright [15, Chapter 6]), we have

$$\begin{aligned} \|r_b\| &= O(\mu), & \|r_c\| &= O(\mu), \\ x_i &= \Omega(1) \ \ (i \in \mathcal{B}), & x_i &= \Omega(\mu) \ \ (i \in \mathcal{N}), \\ s_i &= \Omega(\mu) \ \ (i \in \mathcal{B}), & s_i &= \Omega(1) \ \ (i \in \mathcal{N}). \end{aligned} \tag{72}$$

These estimates are also observed to hold in practice on the majority of problems for values of μ greater than $\mathbf{u}^{1/2}$. An immediate consequence of these estimates and the definition (15) is that

$$D_{ii}^2 = \Omega(\mu^{-1}) \ \ (i \in \mathcal{B}), \qquad D_{ii}^2 = \Omega(\mu) \ \ (i \in \mathcal{N}). \tag{73}$$

We assume the coefficient matrix A to be well conditioned; that is, $\sigma_1(A)$ and $\sigma_m(A)$ are both $\Omega(1)$. We assume further that the submatrix $A_{\cdot\mathcal{B}}$ of columns $A_{\cdot i}$, $i \in \mathcal{B}$, is well conditioned. It follows from this assumption together with the estimate (73) that the matrix $A_{\cdot\mathcal{B}} D_{\mathcal{B}\mathcal{B}}^2 A_{\cdot\mathcal{B}}^T$ has full rank $\min(|\mathcal{B}|, m)$. In fact, since $A_{\cdot\mathcal{B}}$ is well conditioned, all nonzero singular values of $A_{\cdot\mathcal{B}} D_{\mathcal{B}\mathcal{B}}^2 A_{\cdot\mathcal{B}}^T$ are $\Omega(\mu^{-1})$ in size. Likewise, it follows from (15) and (73) that $A_{\cdot\mathcal{N}} D_{\mathcal{N}\mathcal{N}}^2 A_{\cdot\mathcal{N}}^T = O(\mu)$, so we conclude that

$$\sigma_i(AD^2A^T) \ = \ \Omega(\mu^{-1}), \qquad i = 1, 2, \ldots, \min(m, |\mathcal{B}|), \tag{74a}$$

$$\sigma_i(AD^2A^T) \ = \ O(\mu), \qquad i = \min(m, |\mathcal{B}|) + 1, \ldots, m. \tag{74b}$$

Since the largest diagonal element of AD^2A^T is also $\Omega(\mu^{-1})$, the scaled coefficient matrix for (16a) is

$$\rho AD^2A^T, \qquad \text{where } \rho = \Omega(\mu). \tag{75}$$

For consistency with Section 3, the singular values of the matrix in (75) are denoted by σ_i^2. From this definition together with (74) and (75), we deduce that

$$\begin{aligned} \text{(76a)} \qquad \sigma_i^2 &= \Omega(1), \qquad i = 1, 2, \ldots, \min(m, |\mathcal{B}|), \\ \text{(76b)} \qquad \sigma_i^2 &= O(\mu^2), \qquad i = \min(m, |\mathcal{B}|) + 1, \ldots, m. \end{aligned}$$

Recalling our notation p of Section 3, we have in this case that

$$p = \min(m, |\mathcal{B}|).$$

The exact Cholesky factor L (see Sections 3 and 4) satisfies

$$LL^T = \rho AD^2A^T. \tag{77}$$

Suppose now that Algorithm **modchol** is used to compute the solution of (16a), where the right-hand-side component r_{xs} is set to its affine-scaling value $XS\mathbf{1}$. This process result in a computed solution $\widehat{\Delta\pi}^{\text{aff}}$ for (16a). The remaining step components $\widehat{\Delta s}^{\text{aff}}$ and $\widehat{\Delta x}^{\text{aff}}$ are obtained by substitution into (16b) and (16c), respectively, again in finite-precision arithmetic. Our main tool for analyzing the errors in the computed step is Theorem 4.2.

Consider the exact affine scaling step $(\Delta x^{\text{aff}}, \Delta\pi^{\text{aff}}, \Delta s^{\text{aff}})$. Standard results for infeasible-interior-point methods (see, for example, [15, Theorem 7.5]), together with the conditions (72), imply that

$$\|(\Delta x^{\text{aff}}, \Delta s^{\text{aff}})\| = O(\mu). \tag{78}$$

(This estimate holds only when μ falls below a data-dependent threshold $\epsilon(A, b, c)$ defined by Wright [15, Chapter 3].) From (16b) and (72), we have

$$(AA^T)\Delta\pi = A^T(-r_c - \Delta s) = O(\mu),$$

so it follows from our assumptions about the well conditioning of A that

$$\Delta\pi^{\text{aff}} = O(\mu). \tag{79}$$

We can be more specific about the sizes of the critical components Δx_i^{aff}, $i \in \mathcal{N}$ and Δs_i^{aff}, $i \in \mathcal{B}$. If we multiply the third block row in (7) by $(XS)^{-1}$ and use the definition (9), we obtain

$$\frac{\Delta x_i^{\text{aff}}}{x_i} + \frac{\Delta s_i^{\text{aff}}}{s_i} = -1, \qquad i = 1, 2, \ldots, n.$$

Therefore, from (72) and (78), we have for $i \in \mathcal{N}$ that

$$\frac{\Delta x_i^{\text{aff}}}{x_i} = -1 + \frac{O(\mu)}{\Omega(1)} = -1 + O(\mu),$$

and therefore, using (72) again, we have

$$\Delta x_i^{\mathrm{aff}} = -x_i + O(\mu^2), \qquad i \in \mathcal{N}. \tag{80}$$

In a similar way, we obtain

$$\Delta s_i^{\mathrm{aff}} = -s_i + O(\mu^2), \qquad i \in \mathcal{B}. \tag{81}$$

From the estimates (78), (80), and (81), we can show that a near-unit step can be taken along the direction $(\Delta x^{\mathrm{aff}}, \Delta \pi^{\mathrm{aff}}, \Delta s^{\mathrm{aff}})$ without violating positivity of the x and s components. Substituting $(\Delta x, \Delta \pi, \Delta s) = (\Delta x^{\mathrm{aff}}, \Delta \pi^{\mathrm{aff}}, \Delta s^{\mathrm{aff}})$ in (14), we have

$$1 - \alpha_{\max} = O(\mu). \tag{82}$$

To verify this estimate, suppose that $s_i + \alpha \Delta s_i^{\mathrm{aff}} = 0$ for some index $i \in \mathcal{B}$. From (81), we have

$$s_i(1-\alpha) + O(\mu^2) = 0,$$

so it follows from (72) that

$$1 - \alpha = O(\mu^2)/s_i = O(\mu).$$

For the corresponding component x_i, we have from (72) and (78) that $x_i = \Omega(1)$ and $\Delta x_i^{\mathrm{aff}} = O(\mu)$. Hence, for all μ sufficiently small and all $\alpha \in [0,1]$, we have $x_i + \alpha \Delta x_i^{\mathrm{aff}} > 0$. Similar logic can be applied to the remaining indices $i \in \mathcal{N}$, thereby completing our verification of (82).

Returning to the *computed* affine-scaling step $(\widehat{\Delta x}^{\mathrm{aff}}, \widehat{\Delta \pi}^{\mathrm{aff}}, \widehat{\Delta s}^{\mathrm{aff}})$, we now apply Theorem 4.2 after checking that its assumptions of are satisfied for small enough μ and reasonable values of $\mathbf{u}$ and ϵ. For double-precision computations, we have $\mathbf{u} \approx 10^{-14}$. Hence, since A is well conditioned, we can expect the condition (62) to hold in all nonpathological circumstances. Because of (76), our assumption (40a) on the singular value distribution clearly holds for all sufficiently small μ. The condition (40b) is satisfied for any reasonable choice of ϵ. The assumption that Algorithm **modchol** correctly identifies the numerical rank (that is, $|\bar{\mathcal{J}}| = p$) is, as we discussed in Section 3, difficult to guarantee, but it was observed to hold on all problems that we tested. The assumption that rounding errors do not interfere with the makeup of the small pivot index set $\mathcal{J}$ is likewise impossible to verify rigorously; but, as discussed in Section 4, it can reasonably be expected to hold when $\epsilon \geq \mathbf{u}$ (64).

A good choice for ϵ—one that satisfies the assumptions just mentioned while keeping the bound (71) as small as possible—is therefore

$$\epsilon = \mathbf{u}. \tag{83}$$

For generality, we continue to use ϵ and $\bar{\epsilon}$ in the analysis that follows, substituting the specific value (83) only at the end.

Having verified that we can reasonably expect Theorem 4.2 to hold for the system (16a), we now estimate the quantities on the right-hand side of (71). From (76a), we have $\sigma_1/\sigma_p = O(1)$, while from (76b), we have $\sigma_{p+1} = O(\mu)$. The general estimate (34) yields $\sigma_1 = O(1)$, while the definition of γ_{m+1} gives the estimate $\gamma_{m+1} = O(\mathbf{u})$.

We need to account, too, for the errors incurred in evaluating the right-hand side of (16a). The floating-point error in forming $r_{xs} = XS\mathbf{1}$ is only $O(\mu\mathbf{u})$ in magnitude, since just a single floating-point multiplication is needed to calculate each component $x_i s_i$ of this vector, and each such element is $O(\mu)$ (see (72)). The residuals r_b and r_c have magnitude $O(\mu)$ in exact arithmetic (see (72)), but they are calculated as differences of $O(1)$ quantities and so contain evaluation error of absolute magnitude $O(\mathbf{u})$. Specifically, componentwise errors in the computed version of r_c are bounded by $\left(|A|^T|\pi| + |s| + |c|\right)\mathbf{u}$, and similarly for r_b. Because of the estimate (73), the errors in r_c are magnified to $(\mu^{-1}\mathbf{u})$ when we multiply by AD^2 in (16a). In fact, this term is the dominant one in the total right-hand-side evaluation error. The errors that occur when we perform floating-point addition of the terms r_b, $AD^2 r_c$, and $AS^{-1}r_{xs}$ are less significant; they lead to additional terms of sizes $O(\mathbf{u})$ and $O(\mu^{-1}\mathbf{u}^2)$. In summary, the total right-hand-side evaluation error is $O(\mu^{-1}\mathbf{u})$. Hence, after scaling by the factor ρ defined in (75), we have

$$\|e\| = O(\mathbf{u}), \tag{84}$$

where e is the error vector of Section 4.

Substituting the estimates (76), (79), and (84) into (71), we have

$$\|L^T(\widehat{\Delta\pi}^{\text{aff}} - \Delta\pi^{\text{aff}})\| \leq \left\{O(\bar{\epsilon}^{1/2}) + O(\mu) + \tau^2 O(\mathbf{u})\right\}\tau O(\mu) + \tau^2 O(\mathbf{u}).$$

If

$$\tau = O(1) \tag{85}$$

(a reasonable estimate when the Cholesky factorization correctly identifies the numerical rank and $A_{\cdot\mathcal{B}}$ is well conditioned), the error bound above simplifies to

$$\|L^T(\widehat{\Delta\pi}^{\text{aff}} - \Delta\pi^{\text{aff}})\| \leq O(\bar{\epsilon}^{1/2}\mu + \mu^2 + \mathbf{u}). \tag{86}$$

From (77) we have that

$$\rho^{1/2}DA^T = QL^T,$$

for some orthogonal matrix Q. Since orthogonal transformations do not affect the Euclidean norm of a vector, we can substitute $\rho^{1/2}DA^T$ for L^T in (86) and use (75) to write

$$\begin{aligned} \|DA^T(\widehat{\Delta\pi}^{\text{aff}} - \Delta\pi^{\text{aff}})\| &= \rho^{-1/2}\|L^T(\widehat{\Delta\pi}^{\text{aff}} - \Delta\pi^{\text{aff}})\| \\ &\leq O(\bar{\epsilon}^{1/2}\mu^{1/2} + \mu^{3/2} + \mu^{-1/2}\mathbf{u}). \end{aligned} \tag{87}$$

Note too that from (58), (65), (79), and (84), we have

$$\|\widehat{\Delta\pi}^{\text{aff}}\| \leq \|\widehat{\Delta\pi}^{\text{aff}} - \tilde{\Delta\pi}^{\text{aff}}\| + \|\tilde{\Delta\pi}^{\text{aff}} - \Delta\pi^{\text{aff}}\| + \|\Delta\pi^{\text{aff}}\| = O(\mu + \mathbf{u}), \tag{88}$$

where $\tilde{\Delta\pi}^{\text{aff}}$ is the approximate solution that would be obtained by Algorithm **modchol** if it was used to solve (16a) in exact arithmetic.

Next, we examine the effect of the error in $\widehat{\Delta\pi}^{\text{aff}}$ and the evaluation error in the right-hand side of (16b) on the calculated step $\widehat{\Delta s}^{\text{aff}}$. From (79) and (88), we have that

$$\|\Delta\pi^{\text{aff}} - \widehat{\Delta\pi}^{\text{aff}}\| \leq \|\Delta\pi^{\text{aff}}\| + \|\widehat{\Delta\pi}^{\text{aff}}\| = O(\mu + \mathbf{u}). \tag{89}$$

Hence, taking into account the $O(\mathbf{u})$ evaluation error in the term r_c, we have immediately from (16b) that

$$\Delta s^{\rm aff} - \widehat{\Delta s}^{\rm aff} = O(\mathbf{u}) - A^T(\Delta\pi^{\rm aff} - \widehat{\Delta\pi}^{\rm aff}) = O(\mu + \mathbf{u}). \tag{90}$$

Clearly, for the "large" components of s—namely, the $i \in \mathcal{N}$ components—errors of this magnitude do not affect the step length $\alpha_{\max}$ to the boundary defined in (14). However, for the critical components $i \in \mathcal{B}$, the estimate (90) is not good enough to guarantee that $\alpha_{\max}$ is close to 1. (Repeating the argument that follows (82), we find only that $1 - \alpha_{\max} = O(1)$.) Fortunately, a refined estimate of the error in the $\mathcal{B}$ components is available. As in (90), we have

$$\Delta s^{\rm aff} - \widehat{\Delta s}^{\rm aff} = -A^T(\Delta\pi^{\rm aff} - \widehat{\Delta\pi}^{\rm aff}) + O(\mathbf{u}) = D^{-1}v + O(\mathbf{u}), \tag{91}$$

where from (87) we have

$$v = DA^T(\widehat{\Delta\pi}^{\rm aff} - \Delta\pi^{\rm aff}) = O(\bar\epsilon^{1/2}\mu^{1/2} + \mu^{3/2} + \mu^{-1/2}\mathbf{u}). \tag{92}$$

From (73), we have $D_{ii} = \Omega(\mu^{-1/2})$ for $i \in \mathcal{B}$, so from (91) we obtain

$$\widehat{\Delta s}_i^{\rm aff} - \Delta s_i^{\rm aff} = O(\bar\epsilon^{1/2}\mu + \mu^2 + \mathbf{u}), \qquad i \in \mathcal{B}. \tag{93}$$

As in the discussion following (82), we find that $s_i + \alpha\widehat{\Delta s}_i^{\rm aff} = 0$ is possible only if

$$1 - \alpha = O(\bar\epsilon^{1/2} + \mu + \mu^{-1}\mathbf{u}). \tag{94}$$

This estimate suggests that near-unit steps can be taken, at least in the $\widehat{\Delta s}^{\rm aff}$ components, provided that μ is significantly larger that $\mathbf{u}$. When $\mu = O(\mathbf{u})$, all bets are off!

Finally, we estimate the errors in the computed version of $\Delta x^{\rm aff}$ (obtained from (16c)) and estimate their effect on the $\alpha_{\max}$. Again, we consider the components $i \in \mathcal{B}$ and $i \in \mathcal{N}$ separately.

For $i \in \mathcal{B}$, the $O(\mu\mathbf{u})$ evaluation error in $(r_{xs})_i$ is magnified by the term $s_i^{-1} = \Omega(\mu^{-1})$. From (93), replacement of $\Delta s^{\rm aff}$ by $\widehat{\Delta s}^{\rm aff}$ yields an additional error of size $O(\bar\epsilon^{1/2}\mu) + O(\mu^2) + O(\mathbf{u})$, which is also magnified by the $\Omega(\mu^{-1})$ factor. The other arithmetic errors are less significant. In summary, we find that

$$\widehat{\Delta x}_i^{\rm aff} - \Delta x_i^{\rm aff} = O(\bar\epsilon^{1/2} + \mu + \mu^{-1}\mathbf{u}), \qquad i \in \mathcal{B}. \tag{95}$$

By the usual reasoning, we find that $x_i + \alpha\widehat{\Delta x}_i^{\rm aff} = 0$ is possible for $i \in \mathcal{B}$ only for α satisfying (94).

For $i \in \mathcal{N}$, the $O(\mu\mathbf{u})$ evaluation error in $(r_{xs})_i$ is not magnified appreciably by s_i^{-1}, while from (90), the $O(\mu + \mathbf{u})$ error in $\Delta s^{\rm aff}$ is actually diminished after multiplication by $s_i^{-1}x_i = O(\mu)$. We find that

$$\widehat{\Delta x}_i^{\rm aff} - \Delta x_i^{\rm aff} = O(\mu\mathbf{u} + \mu^2), \qquad i \in \mathcal{N}. \tag{96}$$

Hence, we can have $x_i + \alpha\widehat{\Delta x}_i^{\rm aff} = 0$ for $i \in \mathcal{N}$ only if

$$|1 - \alpha| = O(\mathbf{u} + \mu). \tag{97}$$

From (94) and (97), we conclude that the value of $\alpha_{\max}$ defined by (14), with the calculated direction $(\widehat{\Delta x}^{\text{aff}}, \widehat{\Delta \pi}^{\text{aff}}, \widehat{\Delta s}^{\text{aff}})$ replacing the exact search direction, satisfies the estimate

$$1 - \alpha_{\max} = O(\bar{\epsilon}^{1/2} + \mu + \mu^{-1}\mathbf{u}). \tag{98}$$

Note from (89), (90), and (96) that, in an *absolute* sense, the errors in $\widehat{\Delta \pi}^{\text{aff}}$, $\widehat{\Delta s}^{\text{aff}}$, and $\widehat{\Delta x_i}^{\text{aff}}$, $i \in \mathcal{N}$ are small. By contrast, the $O(\mu^{-1}\mathbf{u})$ term in (95) implies that the errors in $\widehat{\Delta x_i}^{\text{aff}}$, $i \in \mathcal{B}$, may become large as $\mu \downarrow 0$. These large errors may in turn cause the residuals r_b to grow as $\mu \downarrow 0$. These expectations are confirmed by the computational experiments of Section 6.

The estimate (98) and the parameter choice $\epsilon = \mathbf{u}$ (83) suggest strongly that the algorithm should be terminated when

$$\mu \le \mathbf{u}^{1/2}. \tag{99}$$

When μ reaches this threshold, all three terms in the estimate (98) are in balance. Below this threshold, the $O(\mu^{-1}\mathbf{u})$ term in $\widehat{\Delta x_i}^{\text{aff}}$ may cause r_b to grow, making further reduction of μ counterproductive. The convergence tolerances used by most interior-point codes—arrived at by practical experience rather than any theoretical considerations—are similar to (99). The code PCx is typical. It declares optimality if the following three conditions are satisfied:

$$\frac{\|r_b\|}{1+\|b\|} \le \texttt{tol}, \quad \frac{\|r_c\|}{1+\|c\|} \le \texttt{tol}, \quad \frac{|c^T x - b^T \pi|}{1+|c^T x|} \le \texttt{tol},$$

where the default value of `tol` is 10^{-8}. (Note that $10^{-8} \approx \mathbf{u}^{1/2}$ in double precision arithmetic on most machines.)

5.2. Mehrotra Predictor-Corrector Steps. Having analyzed the affine-scaling search direction and its calculated approximation, we turn our attention briefly to the search direction used by Mehrotra's predictor-corrector algorithm. As mentioned in Section 2, these steps are obtained by setting r_{xs} as in (12), for some heuristic choice of the centering parameter ζ. We can write the search direction as

$$(\Delta x, \Delta \pi, \Delta s) = (\Delta x^{\text{aff}}, \Delta \pi^{\text{aff}}, \Delta s^{\text{aff}}) + (\Delta x^{\text{cc}}, \Delta \pi^{\text{cc}}, \Delta s^{\text{cc}}), \tag{100}$$

where $(\Delta x^{\text{cc}}, \Delta \pi^{\text{cc}}, \Delta s^{\text{cc}})$ is the "corrector-centering" step component that satisfies the following linear system:

$$\begin{bmatrix} 0 & A^T & I \\ A & 0 & 0 \\ S & 0 & X \end{bmatrix} \begin{bmatrix} \Delta x^{\text{cc}} \\ \Delta \pi^{\text{cc}} \\ \Delta s^{\text{cc}} \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \\ \zeta\mu\mathbf{1} - \Delta X^{\text{aff}} \Delta S^{\text{aff}} \mathbf{1} \end{bmatrix}.$$

Block elimination on this system yields the following special case of (16a):

$$AD^2A^T \Delta \pi^{\text{cc}} = AD^2 \left[X^{-1}(\zeta\mu\mathbf{1} - \Delta X^{\text{aff}} \Delta S^{\text{aff}} \mathbf{1}) \right].$$

Since we assume full rank of A, and since the diagonal elements of D are all strictly positive, the coefficient matrix is invertible, and we have

$$\|\Delta \pi^{\text{cc}}\| \le \|(AD^2A^T)^{-1}AD^2\| \, \|X^{-1}\| \, \|\zeta\mu\mathbf{1} - \Delta X^{\text{aff}} \Delta S^{\text{aff}} \mathbf{1}\|.$$

A result of Stewart [9] and Todd [11] states that the norm $\|(AD^2A^T)^{-1}AD^2\|$ is bounded independently of D over the set of all positive definite diagonal matrices D (and therefore independently of x and s with $(x,s) > 0$). Therefore, we have

$$\|\Delta\pi^{\mathrm{cc}}\| = O(\|X^{-1}\|)O(\|\zeta\mu\mathbf{1} - \Delta X^{\mathrm{aff}}\Delta S^{\mathrm{aff}}\mathbf{1}\|).$$

From (72), we have $\|X^{-1}\| = O(\mu^{-1})$, while from (78), it follows that $\|\Delta X^{\mathrm{aff}}\Delta S^{\mathrm{aff}}\mathbf{1}\| = O(\mu^2)$. Hence, we have

$$\|\Delta\pi^{\mathrm{cc}}\| = O(\zeta + \mu). \tag{101}$$

A typical heuristic for choosing the centering parameter ζ is to set

$$\zeta = (\mu_{\mathrm{aff}}/\mu)^3,$$

where μ_{aff} is the value of μ that results from a full step-to-boundary $\alpha_{\max}$ along the affine-scaling direction. If the search direction is exact, we have $\mu_{\mathrm{aff}} = O(\mu^2)$, so this heuristic yields $\zeta = O(\mu^3)$. Use of the calculated direction $(\widehat{\Delta x}^{\mathrm{aff}}, \widehat{\Delta\pi}^{\mathrm{aff}}, \widehat{\Delta s}^{\mathrm{aff}})$ together with the estimate (98) leads us to expect $\mu_{\mathrm{aff}} = O(\mu^2)$ in this case too, provided that $\mu \geq \mathbf{u}^{1/2}$. Hence, we have from (101) that $\|\Delta\pi^{\mathrm{cc}}\| = O(\mu)$, and so, from (100) and (79), we have

$$\|\Delta\pi\| = O(\mu), \tag{102}$$

where $\Delta\pi$ is the π component of the Mehrotra search direction.

We also can apply the Stewart-Todd result to formulae for Δx^{cc} and Δs^{cc} to show that $\|(\Delta x^{\mathrm{cc}}, \Delta s^{\mathrm{cc}})\| = O(\mu)$. Therefore, we have

$$\|(\Delta x, \Delta s)\| = O(\mu), \tag{103}$$

corresponding to (78).

Because of the estimates (102) and (103), the analysis of the preceding subsection can be applied without modification to the calculated version of the search direction (100). In particular, if we redefine the step-to-boundary $\alpha_{\max}$ in terms of this calculated step $(\widehat{\Delta x}, \widehat{\Delta\pi}, \widehat{\Delta s})$, we find that the estimate (98) still applies. We conclude that near-unit steps can still be taken along this direction provided that $\mu \geq \mathbf{u}^{1/2}$.

6. Implementation and Computational Results. Most interior-point codes use modified Cholesky algorithms with essentially the same properties as Algorithm **modchol**. They differ slightly, however, in the implementation. The IPMOS code of Xu, Hung, and Ye [16] replaces small pivot elements by 1 and fills out the corresponding column of the Cholesky factor with zeros and also inserts a zero in the right-hand side. The criterion for identifying a small pivot is not explained in the reference [16], but otherwise this strategy is equivalent to Algorithm **modchol**. Zhang's LIPSOL code [17] and the PCx code of Czyzyk, Mehrotra, and Wright [1] replace small pivots by a huge number—10^{128}—but otherwise leave the Cholesky algorithm unchanged. The net effect is, however, almost equivalent to Algorithm **modchol** and the triangular substitution procedure (25). The advantage of this approach is that it involves minimal changes to a standard sparse Cholesky code. We need only add a loop to calculate the largest diagonal element β, and a small pivot check immediately before the point at which the computation $L_{ii} = \sqrt{M_{ii}}$ is performed.

To test that the analysis of this paper was reflected in practical computations, we coded a primal-dual algorithm that used Algorithm **modchol** in conjunction with the formulation (16). The code was used to solve some small random linear programs in which the amount of degeneracy—the composition of index sets $\mathcal{B}$ and $\mathcal{N}$—was carefully controlled. At each iterate, we monitored various quantities and compared them against the estimates of Section 5.

The linear programming test problems were posed in standard form (2) with $m = 6$ and $n = 12$. The matrix A is fully dense, with elements $(\tau_1 - .5)10^{6(\tau_2-.5)}$, where τ_1 and τ_2 are random variables drawn from a uniform distribution on the interval $[0, 1]$. (Of course, the values of τ_1 and τ_2 are different for each element of the matrix.) We can reasonably expect this matrix A to satisfy the well-conditioning assumptions of Section 5. The user specifies the number of indices to appear in $\mathcal{B}$, and we set

$$|\mathcal{N}| = n - |\mathcal{B}|, \qquad \mathcal{N} = \{1, 2, \cdots, |\mathcal{N}|\}, \qquad \mathcal{B} = \{|\mathcal{N}| + 1, \cdots, n\}.$$

A primal solution x^* is constructed with

$$x_i^* = 0 \;\; (i = 1, 2, \cdots, |\mathcal{N}|), \qquad x_i^* = 10^{3(\tau-.5)} \;\; (i = |\mathcal{N}| + 1, \cdots, n),$$

where τ is randomly drawn from the uniform distribution on $[0, 1]$. We choose the dual solution π^* to be the vector $(1, 1, \cdots, 1)^T$, and fix an optimal dual slack vector s^* to be

$$s_i^* = 10^{4(\tau-.5)} \;\; (i = 1, 2, \cdots, |\mathcal{N}|), \qquad s_i^* = 0 \;\; (i = |\mathcal{N}| + 1, \cdots, n),$$

where τ is random as above. Finally, we set $b = Ax^*$ and $c = A^T\pi^* + s^*$.

The code was an implementation of the infeasible-interior-point algorithm described by Wright [13]. The details of this algorithm are unimportant; we need note only that its iterates satisfy the estimates (72) in exact arithmetic and that the algorithm takes steps along the affine scaling direction during its later iterations. At each iteration of the algorithm, we calculated the affine scaling direction (whether or not it was actually used as a search direction) and printed the norms $\|\widehat{\Delta x}^{\text{aff}}\|_\infty$, $\|\widehat{\Delta \pi}^{\text{aff}}\|_\infty$, and $\|\widehat{\Delta s}^{\text{aff}}\|_\infty$ alongside the duality measure μ and residual norm $\|(r_b, r_c)\|_\infty$ for the current point. We also kept track of the number of small pivots encountered during the factorization, that is, the number of elements in $\mathcal{J}$. The parameter ϵ was set to 10^{-12}, which is about $100\mathbf{u}$ on the SPARCstation 5 that was used for the experiments. The results were not particularly sensitive to this parameter.

Results are shown in Tables 1–4. For each iteration of the algorithm, these tables list the number of small pivots $|\mathcal{J}|$, the base-10 logarithms of μ, $\|(r_b, r_c)\|_\infty$, and the affine-scaling step norms mentioned above. The step-to-boundary $\alpha_{\max}$ along the calculated affine-scaling direction is also tabulated. A horizontal line in each table indicates the iterate at which termination occurs according to the criterion (99).

In Table 1 we chose $|\mathcal{B}| = m = 6$, making the linear program nondegenerate and the primal-dual solution unique. It is clear that $\widehat{\Delta \pi}^{\text{aff}}$ and $\widehat{\Delta s}^{\text{aff}}$ satisfy the estimates (88) and (90), respectively, even when the algorithm is continues past the point of normal termination. The component $\widehat{\Delta x}^{\text{aff}}$, on the other hand, clearly shows the influence of the $O(\mu^{-1}\mathbf{u})$ error term in (95) when μ becomes comparable to or smaller than $\mathbf{u}$. Note, too, that the error in $\widehat{\Delta x}^{\text{aff}}$ is transmitted to the residual r_b

on succeeding iterations but that this effect does not become destructive until μ is much smaller than its normal termination threshold. The values of $\alpha_{\max}$ are also consistent with the estimate (98). This step length approaches 1 until the normal point of termination is reached, after which the errors in $\widehat{\Delta x}^{\text{aff}}$ and r_b make further progress impossible.

Table 2 shows the interesting case in which we choose $|\mathcal{B}| = 4$, so that the coefficient matrix in (16a) has four singular values of magnitude $\Omega(\mu^{-1})$ and two of magnitude $\Omega(\mu)$. The second column shows that Algorithm **modchol** correctly identifies the numerical rank during the last few iterations and that the interior-point algorithm continues to generate useful steps and to make good progress even after **modchol** encounters small pivots. Apart from this feature, the behavior is the same as in Table 1, with errors in $\widehat{\Delta x}^{\text{aff}}$ causing the interior-point algorithm to behave poorly when it is permitted to run past its normal point of termination. We noted that for all iterations, the "small" pivots were at the bottom right corner of the Cholesky matrix, so that (28) rather than the general estimate (27) applies to the perturbation matrix E. In this case, we can replace $\bar{\epsilon}^{1/2}$ by $\bar{\epsilon}$ in estimates of Section 5 such as (93), (95), and (98).

Table 3 illustrates another case in which $|\mathcal{B}| = 4$, with the added complication that A is rank deficient. (We forced rank deficiency by setting $A_{1j} = 0$ and $A_{2j} = 0$ for $j = 1, 2, \cdots, n-1$, so that the first and second rows each contain a single nonzero in their last column.) The $(2, 2)$ pivot is skipped at every invocation of Algorithm *modchol*. As μ becomes small, the final pivot is skipped as well, and the numerical rank is correctly determined. Since the small pivots are not localized in the bottom right corner, the special bound (28) does not apply, so we cannot strengthen the bounds on the step components as in the previous paragraph. The computational behavior is qualitatively the same as in Tables 1 and 2.

Table 4 illustrates a problem for which $|\mathcal{B}| = 8$. Here, the coefficient matrices retain full numerical rank at all iterates, and the behavior is similar to that reported in Table 1. One point of difference is that the errors in $\widehat{\Delta x}^{\text{aff}}$, which start to increase after iteration 19, do not have an immediate effect on the residual r_b. The reason is simply that this particular interior-point algorithm chose to take a path-following step at iterations 21 and 22 rather than the affine scaling step, and the Δx components were calculated accurately in the path following step. An affine-scaling step is, however, taken at iteration 28, and the effect of the error in $\widehat{\Delta x}^{\text{aff}}$ on the residual r_b at the following iterate is obvious.

REFERENCES

[1] J. Czyzyk, S. Mehrotra, and S. J. Wright, *PCx User Guide*, Technical Report OTC 96/01, Optimization Technology Center, Argonne National Laboratory and Northwestern University, May 1996.

[2] A. Forsgren, P. Gill, and J. Shinnerl, *Stability of symmetric ill-conditioned systems arising in interior methods for constrained optimization*, SIAM Journal on Matrix Analysis and Applications, 17 (1996), pp. 187–211.

[3] A. George and J. W.-H. Liu, *Computer Solution of Large Sparse Positive Definite Systems*, Prentice-Hall, 1981.

[4] N. J. Higham, *Accuracy and Stability of Numerical Algorithms*, SIAM Publications, Philadelphia, 1996.

[5] C. L. Lawson and R. J. Hanson, *Solving Least Squares Problems*, Prentice-Hall, Englewood Cliffs, NJ, 1974.

Table 1

Affine scaling step characteristics for a problem with $m = 6$, $n = 12$, $|\mathcal{B}| = 6$. $\|\cdot\| = \|\cdot\|_\infty$, and the horizontal line represents the normal point of termination.

	Small		log	log	log	log	
Iteration	Pivots	$\log \mu$	$\|(r_b, r_c)\|$	$\|\widehat{\Delta x}^{\text{aff}}\|$	$\|\widehat{\Delta \pi}^{\text{aff}}\|$	$\|\widehat{\Delta s}^{\text{aff}}\|$	$\alpha_{\max}$
⋮							
12	0	-0.6	-11.1	-0.1	-0.6	0.6	.26426
13	0	-1.4	-10.7	0.4	-1.1	0.1	.77520
14	0	-2.1	-10.7	1.2	-2.3	-1.1	.39373
15	0	-3.3	-10.4	-0.3	-1.3	-0.1	.62276
16	0	-4.8	-8.1	-1.1	-5.2	-3.9	.99697
17	0	-7.2	-10.5	-3.5	-8.3	-7.1	.99999
18	0	-12.0	-12.2	-8.2	-14.0	-12.5	>.99999
19	0	-21.0	-12.0	-3.6	-14.9	-13.9	.99975
20	0	-24.2	-4.6	-1.4	-15.0	-13.9	.93989
21	0	-26.2	-1.5	1.4	-15.3	-14.5	.06843
⋮							

Table 2

Affine scaling step characteristics for a problem with $m = 6$, $n = 12$, $|\mathcal{B}| = 4$. $\|\cdot\| = \|\cdot\|_\infty$, and the horizontal line represents the normal point of termination.

	Small		log	log	log	log	
Iteration	Pivots	$\log \mu$	$\|(r_b, r_c)\|$	$\|\widehat{\Delta x}^{\text{aff}}\|$	$\|\widehat{\Delta \pi}^{\text{aff}}\|$	$\|\widehat{\Delta s}^{\text{aff}}\|$	$\alpha_{\max}$
⋮							
12	0	-0.6	-12.0	0.1	-1.3	0.7	.95133
13	0	-1.9	-11.4	-1.5	-0.2	1.8	.51719
14	0	-2.4	-9.5	-1.8	-0.9	1.0	.90453
15	1	-3.4	-9.3	-2.7	-5.5	-3.5	.98770
16	2	-5.2	-9.1	-4.4	-7.2	-5.2	.99977
17	2	-8.5	-11.1	-7.7	-10.5	-8.5	>.99999
18	2	-14.4	-13.2	-12.5	-15.9	-13.8	>.99999
19	2	-25.1	-12.3	-2.1	-15.9	-13.7	>.99999
20	2	-30.4	-1.8	5.0	-15.9	-14.5	.00016
21	3	-30.4	2.3	10.6	-16.1	-13.5	<.00001
⋮							

Table 3

Affine scaling step characteristics for a problem with $m = 6$, $n = 12$, $|\mathcal{B}| = 4$, in which A is rank deficient. $\|\cdot\| = \|\cdot\|_\infty$, and the horizontal line represents the normal point of termination.

	Small		log	log	log	log	
Iteration	Pivots	$\log \mu$	$\|(r_b, r_c)\|$	$\|\widehat{\Delta x}^{\text{aff}}\|$	$\|\widehat{\Delta \pi}^{\text{aff}}\|$	$\|\widehat{\Delta s}^{\text{aff}}\|$	$\alpha_{\max}$
⋮							
12	1	-1.4	-11.3	0.1	0.5	1.0	.78614
13	1	-2.1	-10.5	-1.9	1.7	2.0	.17726
14	1	-2.7	-9.2	-0.8	1.2	1.3	.41306
15	1	-2.9	-9.1	0.1	0.5	0.6	.00442
16	1	-3.2	-8.9	-0.6	0.2	0.4	.78585
17	1	-3.9	-8.8	-1.1	-2.0	-1.7	.93466
18	1	-4.8	-9.7	-2.0	-2.0	-1.7	99179
19	2	-6.2	-10.9	-3.4	-6.0	-5.5	99970
20	2	-8.6	-10.1	-5.8	-8.3	-7.9	$>$.99999
21	2	-12.7	-10.9	-10.2	-12.7	-12.0	$>$.99999
22	2	-20.2	-11.5	-4.1	-12.3	-12.5	.99988
23	2	-21.7	-4.6	-2.6	-12.9	-12.2	$>$.99999
24	2	-27.3	-3.0	2.9	-12.5	-12.3	.00711
⋮							

Table 4

Affine scaling step characteristics for a problem with $m = 6$, $n = 12$, $|\mathcal{B}| = 8$. $\|\cdot\| = \|\cdot\|_\infty$, and the horizontal line represents the normal point of termination.

	Small		log	log	log	log	
Iteration	Pivots	$\log \mu$	$\|(r_b, r_c)\|$	$\|\widehat{\Delta x}^{\text{aff}}\|$	$\|\widehat{\Delta \pi}^{\text{aff}}\|$	$\|\widehat{\Delta s}^{\text{aff}}\|$	$\alpha_{\max}$
⋮							
12	0	-0.1	-5.7	3.5	-3.6	-2.3	.42851
13	0	-0.9	-9.5	2.1	-2.4	-1.4	.60234
14	0	-1.3	-9.9	1.8	-2.8	-1.7	.38898
15	0	-2.0	-10.6	1.8	-1.8	-0.6	.30608
16	0	-2.2	-10.6	1.2	-1.8	-0.6	.37400
17	0	-3.3	-10.5	-1.0	-3.9	-2.7	.68815
18	0	-4.2	-11.3	-0.3	-4.1	-2.9	.99691
19	0	-6.7	-9.7	-3.0	-7.5	-6.3	.99998
20	0	-11.3	-10.6	0.6	-12.4	-11.2	.98674
21	0	-13.2	-10.4	2.3	-14.4	-13.1	.19743
22	0	-13.3	-10.5	2.3	-14.6	-13.5	.04721
⋮							
28	0	-18.8	$-\infty$	-2.2	-18.5	-14.4	.99999
29	0	-23.7	-5.5	1.1	-15.0	-13.7	.59156
⋮							

[6] I. J. Lustig, R. E. Marsten, and D. F. Shanno, *Computational experience with a primal-dual interior point method for linear programming*, Linear Algebra and Its Applications, 152 (1991), pp. 191–222.

[7] C. Mészáros, *The inexact minimum local fill-in ordering algorithm*, Working Paper 95–7, Computer and Automation Institute, Hungarian Academy of Sciences, Budapest, 1995.

[8] E. Ng and B. W. Peyton, *Block sparse Cholesky algorithms on advanced uniprocessor computers*, SIAM Journal on Scientific Computing, 14 (1993), pp. 1034–1056.

[9] G. W. Stewart, *On scaled projections and pseudoinverses*, Linear Algebra and Its Applications, 112 (1989), pp. 189–193.

[10] G. W. Stewart and J. Sun, *Matrix Perturbation Theory*, Computer Science and Scientific Computing, Academic Press, New York, 1990.

[11] M. J. Todd, *A Dantzig-Wolfe-like variant of Karmarkar's interior-point linear programming algorithm*, Operations Research, 38 (1990), pp. 1006–1018.

[12] M. H. Wright, *Some properties of the Hessian of the logarithmic barrier function*, Mathematical Programming, 67 (1994), pp. 265–295.

[13] S. J. Wright, *A path-following interior-point algorithm for linear and quadratic optimization problems*, Preprint MCS–P401–1293, Mathematics and Computer Science Division, Argonne National Laboratory, Argonne, Ill., December 1993. To appear in Annals of Operations Research.

[14] ———, *Stability of augmented system factorizations in interior-point methods*, Preprint MCS–P446–0694, Mathematics and Computer Science Division, Argonne National Laboratory, Argonne, Ill., June 1994. Revised, July 1995. To appear in SIAM Journal of Matrix Analysis and Applications.

[15] ———, *Primal-Dual Interior-Point Methods*, SIAM, Philadelphia, Pa, 1996. To appear.

[16] X. Xu, P. Hung, and Y. Ye, *A simplified homogeneous and self-dual linear programming algorithm and its implementation.* To appear in Annals of Operations Research, September 1993.

[17] Y. Zhang, *Solving large-scale linear programs by interior-point methods under the MATLAB enviroment*, Technical Report TR96-01, Department of Mathematics and Statistics, University of Maryland Baltimore County, Baltimore, Md, 1996.