SIAM J. OPTIM.
Vol. 9, No. 4, pp. 1159–1191

MODIFIED CHOLESKY FACTORIZATIONS IN INTERIOR-POINT ALGORITHMS FOR LINEAR PROGRAMMING*

STEPHEN J. WRIGHT†

To John Dennis, with appreciation, on the occasion of his 60th birthday.

Abstract. We investigate a modified Cholesky algorithm typical of those used in most 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 algorithms (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 alternative factorizations; and they perform robustly on most practical problems, yielding good interior-point steps even when the coefficient matrix of the main linear system to be solved for the step components is ill conditioned. We investigate this surprisingly robust performance by using analytical tools from matrix perturbation theory and error analysis, illustrating our results with computational experiments. Finally, we point out the potential limitations of this approach.

Key words. interior-point algorithms and software, Cholesky factorization, matrix perturbations, error analysis

AMS subject classifications. 65F05, 65G05, 90C05

PII. S1052623496304712

1. Introduction.

Most interior-point codes for linear programming share a common feature: their major computational operation at each iteration—solution of a large system of linear equations with a symmetric positive definite coefficient matrix—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, 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 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 modify the Cholesky procedure so that it skips the unacceptable pivots or replaces them with workable values. For instance, the offending pivot element is sometimes replaced by a huge number, as in LIPSOL [20] and PCx [3]. In other codes, such as IPMOS [19], the pivot is replaced by a moderate number, but the corresponding right-hand-side

*Received by the editors June 4, 1996; accepted for publication (in revised form) November 14, 1998; published electronically September 24, 1999. 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. The U.S. Government retains a nonexclusive, royalty-free license to publish or reproduce the published form of this contribution, or allow others to do so, for U.S. Government purposes. Copyright is owned by SIAM to the extent not limited by these rights.

http://www.siam.org/journals/siopt/9-4/30471.html

†Mathematics and Computer Science Division, Argonne National Laboratory, 9700 South Cass Avenue, Argonne, IL 60439 (wright@mcs.anl.gov).

element is set to zero, as are the off-diagonal elements in the corresponding column of the Cholesky factor. The net effects of these approaches, and the approaches used in other Cholesky-based codes, such as OB1 [9], HOPDM [6], and the APOS code of XPRESS-MP [1], are all quite similar to those of the algorithm **modchol** that we analyze in this paper: each small or negative pivot causes the Cholesky procedure to skip one stage, and the solution component corresponding to this pivot is set to zero (or to a very small number). To date, there has been little investigation of these pivot-skipping strategies from a numerical analysis viewpoint.

In the context of Cholesky factorization of general symmetric positive semidefinite matrices, Lawson and Hanson [8, p. 125] advocated the use of pivot skipping when negative pivots are encountered. They also suggested the alternative remedy of diagonal pivoting, in which a "large" diagonal element is selected from the unreduced portion of the matrix at each stage and moved to the pivot position by a symmetric row and column exchange. The procedure terminates when none of the remaining diagonal elements is large enough to qualify as a pivot, and an approximate solution is computed with the partial factors. Higham [7, Chapter 10] presented an error analysis of this approach, and M. H. Wright [15] considered its use in factoring the Hessian matrices that arise in the Newton/logarithmic-barrier method for nonlinear programming. This strategy is not practical in the context of interior-point linear programming codes because the matrices in question are too large to allow row and column exchanges to be performed efficiently. On the other hand, pivot-skipping strategies have the advantage that they can be implemented by changing just a few lines of a general sparse Cholesky code, so it is 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 [20] and PCx [3] make explicit use of the sparse Cholesky code of Ng and Peyton [10].) Moreover, the good practical performance of pivot-skipping strategies made the search for alternatives less urgent.

In this paper, we investigate the good performance of pivot-skipping strategies on the majority of practical problems. In section 3, we specify our representative pivot-skipping strategy, which we term **modchol** for convenience, and analyze the effects of the skipped pivots on the computed triangular factor and computed solution. In section 4, we incorporate the effects of finite-precision arithmetic into the analysis. Both sections are general in that they apply to general symmetric positive semidefinite matrices, not just the specific matrices that arise in the interior-point application. In section 5, however, we apply the results of sections 3 and 4 to the equations for calculating the interior-point step, showing how the errors in the computed steps affect the progress of the interior-point algorithm, suggesting a suitable termination criterion, and indicating possible shortcomings in the pivot-skipping approach. Our analysis in this section applies to primal- and dual-degenerate linear programs. We conclude with some computational results in section 6.

A number of other theoretical papers on linear algebra operations in barrier and interior-point methods have appeared in recent years. We mentioned above the paper of M. H. Wright [15], in which a Cholesky procedure with diagonal pivoting was used as the basis of an algorithm to construct steps that are accurate both in the subspace spanned by the active constraint Jacobian and its complement. Our focus in the current paper is on (possibly degenerate) linear programs rather than nondegenerate nonlinear programs. Moreover, we do not allow diagonal pivoting and, since our problem is a linear program, the issue of resolving the component of the step in the near-null space of the active constraint matrix is not as relevant.

In an earlier paper [18], S. J. Wright considered the stability of algorithms for the symmetric indefinite formulation of the step equations at each iteration of an interior-point method for linear programming. Ill conditioning of the coefficient matrix is the key issue in this formulation as well, but we showed that, in general, the calculated steps are good search directions for the interior-point method. Forsgren, Gill, and Shinnerl [5] perform a similar analysis in the context of logarithmic barrier methods for nonlinear problems, but they assume a certain ordering of the rows and columns of the coefficient matrix.

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

The ith singular value of a matrix B is denoted by $\sigma_i(B)$. 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 sets $\mathcal{I}$ and $\mathcal{K}$, $M_{\mathcal{I}\mathcal{K}}$ denotes the submatrix formed by the elements M_{ij} for $i \in \mathcal{I}$ and $j \in \mathcal{K}$. The jth column of M is denoted by $M_{\cdot j}$, the column submatrix consisting of columns $j \in \mathcal{K}$ is denoted by $M_{\cdot\mathcal{K}}$, and the submatrix of elements M_{ij} for $j \in \mathcal{K}$ is noted by $M_{i,\mathcal{K}}$. The submatrix consisting of rows and columns i through j is denoted by $M_{i:j,i:j}$.

Unit roundoff error, which we denote by $\mathbf{u}$, can be defined implicitly by the following statement (see, for example, Higham [7]). When α and ζ are any two floating-point numbers; op denotes $+$, $-$, $\times$, and $/$; and $\mathrm{fl}(\cdot)$ denotes the floating-point representation of a real number, we have

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

We use $\mathrm{comp}(\cdot)$ to denote the calculated version of the quantity in question, taking into account the effects of roundoff error.

In estimating the sizes of various quantities that arise in the analysis, we use δ_1 to denote a constant whose magnitude depends at most cubically on the dimension m of the linear system. We often use $\delta_{\mathbf{u}}$ as a shorthand for $\delta_1 \mathbf{u}$. Order notation $O(\cdot)$ and $\Theta(\cdot)$ is used as follows: if v (vector or scalar) and ϵ (nonnegative scalar) are two quantities that share a dependence on other variables, we write $v = O(\epsilon)$ if there is a moderate constant β_1 such that $\|v\| \le \beta_1 \epsilon$ for all values of ϵ that are either sufficiently close to zero or sufficiently large, depending on the context. We write $v = \Theta(\epsilon)$ if there are constants β_1 and β_0 such that $\beta_0 \epsilon \le \|v\| \le \beta_1 \epsilon$ for ϵ in the ranges specified above.

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 $(1, 1, \ldots, 1)^T$.

Finally, we mention the parameter ϵ that defines the pivot threshold in the modified Cholesky algorithm. A scaled quantity $\bar{\epsilon}$ defined by

$$\bar{\epsilon} \stackrel{\text{def}}{=} 2m^2 \epsilon \tag{1.1}$$

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

2. Primal-dual algorithms for linear programming.

We consider the linear programming problem in standard form:

$$\min\, c^T x \quad \text{subject to} \quad Ax = b, \qquad x \ge 0, \tag{2.1}$$

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 (2.1) is

$$\max\, b^T \pi \quad \text{subject to} \quad A^T \pi + s = c, \qquad s \ge 0, \tag{2.2}$$

where $s \in \mathsf{R}^n$ and $\pi \in \mathsf{R}^m$. We assume throughout the paper that A has full row rank (which can be guaranteed by preprocessing the data), so that $m \leq n$. The Karush–Kuhn–Tucker (KKT) conditions, which identify a vector triple (x, π, s) as a primal-dual solution for (2.1), (2.2), can be stated as follows:

$$A^T\pi + s = c, \tag{2.3a}$$

$$Ax = b, \tag{2.3b}$$

$$x_i s_i = 0, \qquad i = 1, 2, \ldots, n, \tag{2.3c}$$

$$(x, s) \geq 0. \tag{2.3d}$$

We assume throughout the paper that a primal-dual solution exists, but we make no assumptions about uniqueness or nondegeneracy. 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

$$x_i^* = 0 \quad \text{for all} \quad i \in \mathcal{N}, \qquad s_i^* = 0 \quad \text{for all} \quad i \in \mathcal{B}. \tag{2.4}$$

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 that is equivalent to the first three KKT conditions (2.3a), (2.3b), (2.3c), namely,

$$Ax - b = 0, \qquad A^T\pi + s - c = 0, \qquad XS\mathbf{1} = 0, \tag{2.5}$$

where $X = \operatorname{diag}(x_1, x_2, \ldots, x_n)$ and $S = \operatorname{diag}(s_1, s_2, \ldots, s_n)$. In general, the search direction $(\Delta x, \Delta\pi, \Delta s)$ satisfies the following linear system:

$$\begin{bmatrix} 0 & A^T & I \\ A & 0 & 0 \\ S & 0 & X \end{bmatrix} \begin{bmatrix} \Delta x \\ \Delta\pi \\ \Delta s \end{bmatrix} = \begin{bmatrix} -r_c \\ -r_b \\ -r_{xs} \end{bmatrix}, \tag{2.6}$$

where the coefficient matrix is the Jacobian of (2.5) and the right-hand-side components r_b and r_c are defined by

$$r_b = Ax - b, \qquad r_c = A^T\pi + s - c. \tag{2.7}$$

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

$$r_{xs} = XS\mathbf{1}, \tag{2.8}$$

and, in this case, we denote the solution of (2.6) by $(\Delta x^{\text{aff}}, \Delta\pi^{\text{aff}}, \Delta s^{\text{aff}})$. In a path-following method, we have

$$r_{xs} = XS\mathbf{1} - \zeta\mu\mathbf{1}, \tag{2.9}$$

where μ is the duality gap defined by

$$\mu = x^T s/n, \tag{2.10}$$

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

$$r_{xs} = XS\mathbf{1} + \Delta X^{\text{aff}}\Delta S^{\text{aff}}\mathbf{1} - \zeta\mu\mathbf{1}, \tag{2.11}$$

where ΔX^{aff} and ΔS^{aff} are the diagonal matrices formed from the affine-scaling step components Δx^{aff} and Δs^{aff}, and the value of ζ is determined by a heuristic based on the effectiveness of the affine-scaling direction. Mehrotra's method requires the solution of *two* linear systems at each iteration—the affine-scaling system (2.6), (2.7), (2.8), and the search direction system (2.6), (2.7), (2.11)—though the coefficient matrix is the same for both systems. Gondzio's [6] higher-order corrector method refines the step by solving additional linear systems, all with the same coefficient matrix as in (2.6).

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,

$$(x, s) + \alpha(\Delta x, \Delta s) > 0. \tag{2.12}$$

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

$$\alpha_{\max} = \sup_{\alpha \in [0,1]} \{\alpha \,|\, (x, s) + \alpha(\Delta x, \Delta s) \geq 0\}. \tag{2.13}$$

A typical strategy is to set

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

where $\eta \in [.9, 1.0)$ approaches 1 as the iterates approach the solution set.

By applying block elimination to (2.6) and using the notation

$$D^2 = S^{-1}X, \tag{2.14}$$

we obtain the following equivalent system:

$$AD^2A^T\Delta\pi = -r_b - AD^2(r_c - X^{-1}r_{xs}), \tag{2.15a}$$

$$\Delta s = -r_c - A^T\Delta\pi, \tag{2.15b}$$

$$\Delta x = -S^{-1}(r_{xs} + X\Delta s). \tag{2.15c}$$

In many codes, the solution is obtained from just this formulation. A sparse Cholesky factorization, modified to handle small or negative pivots, is applied to the symmetric positive definite coefficient matrix AD^2A^T in (2.15a) and the solution $\Delta\pi$ is obtained by triangular substitution with the computed factor. The remaining direction components are recovered from (2.15b) and (2.15c). Computational experience shows that this technique yields steps that are useful search directions for the interior-point algorithm, even when AD^2A^T is ill conditioned and when the computed version of $\Delta\pi$ has few digits in common with the exact version. 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.

In section 5, we investigate this phenomenon by applying the error analysis developed in sections 3 and 4 to the solution of the system (2.15), assuming that our algorithm **modchol** is used to solve (2.15a) and that all computations are performed in finite-precision floating-point arithmetic. We examine the effects of the errors in the computed step on properties such as the value of $\alpha_{\max}$ (2.13) and on the updated values of the residuals r_b and r_c—properties that indicate whether the step is a useful one for the interior-point method.

We start by specifying **modchol** and analyzing its properties as they pertain to a general linear system $Mz = r$, where M is symmetric positive definite.

3. A modified Cholesky algorithm. In this section, we describe and analyze **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. The output of **modchol** is 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 ϵ 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

$$E^{(i)} = \left[\begin{array}{c|cccc} 0 & 0 & \cdots & \cdots & 0 \\ \hline 0 & M_{ii}^{(i-1)} & \cdots & \cdots & M_{im}^{(i-1)} \\ \vdots & \vdots & 0 & & 0 \\ \vdots & \vdots & & \ddots & \vdots \\ 0 & M_{mi}^{(i-1)} & 0 & \cdots & 0 \end{array}\right], \qquad M^{(i)} = M^{(i-1)} - E^{(i)}; \tag{3.1}$$

  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

$$E = \sum_{i \in \mathcal{J}} E^{(i)} \tag{3.2}$$

and denote the complement of $\mathcal{J}$ in $\{1, 2, \ldots, m\}$ by $\bar{\mathcal{J}}$, it follows from (3.1) that

$$E_{\bar{\mathcal{J}}\bar{\mathcal{J}}} = 0. \tag{3.3}$$

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

$$\tilde{L}\tilde{L}^T = \tilde{M} = M - E. \tag{3.4}$$

By partitioning this equation into its $\mathcal{J}$ and $\bar{\mathcal{J}}$ components and using $\tilde{L}_{\cdot\mathcal{J}} = 0$ and (3.3), we obtain

$$M_{\bar{\mathcal{J}}\bar{\mathcal{J}}} = \tilde{L}_{\bar{\mathcal{J}}\cdot}\tilde{L}_{\bar{\mathcal{J}}\cdot}^T + E_{\bar{\mathcal{J}}\bar{\mathcal{J}}} = \tilde{L}_{\bar{\mathcal{J}}\bar{\mathcal{J}}}\tilde{L}_{\bar{\mathcal{J}}\bar{\mathcal{J}}}^T, \tag{3.5a}$$

$$M_{\bar{\mathcal{J}}\mathcal{J}} = \tilde{L}_{\bar{\mathcal{J}}\cdot}\tilde{L}_{\mathcal{J}\cdot}^T + E_{\bar{\mathcal{J}}\mathcal{J}} = \tilde{L}_{\bar{\mathcal{J}}\bar{\mathcal{J}}}\tilde{L}_{\mathcal{J}\bar{\mathcal{J}}}^T + E_{\bar{\mathcal{J}}\mathcal{J}}. \tag{3.5b}$$

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

$$LL^T = M. \tag{3.6}$$

Given the linear system

$$Mz = r, \tag{3.7}$$

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

$$z = M^{-1}r = L^{-T}L^{-1}r. \tag{3.8}$$

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 $\tilde{z}_{\mathcal{J}}$ is set to zero. From (3.5a), we see that $\tilde{z}_{\bar{\mathcal{J}}}$ can be calculated by performing a pair of triangular substitutions; that is,

$$\tilde{z}_{\bar{\mathcal{J}}} = \tilde{L}_{\bar{\mathcal{J}}\bar{\mathcal{J}}}^{-T}\tilde{L}_{\bar{\mathcal{J}}\bar{\mathcal{J}}}^{-1}r_{\bar{\mathcal{J}}}, \qquad \tilde{z}_{\mathcal{J}} = 0. \tag{3.9}$$

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. However, in the main result of this section (Theorem 3.4), we show that the difference between $\tilde{L}^T z$ and $\tilde{L}^T \tilde{z}$ is small. As we see in section 5, this difference determines the usefulness of the computed solution of (2.15) as a search direction for the interior-point algorithm.

To simplify the analysis, we assume throughout the paper that

$$\beta = 1. \tag{3.10}$$

A trivial scaling, which affects neither the algorithm nor its analysis, can always be applied to the symmetric positive definite matrix M to yield (3.10).

We start with a simple result about the intermediate matrices $M^{(i)}$ that arise during **modchol**.

LEMMA 3.1. *If* (3.10) *holds, then 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 same 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 $M^{(i-1)}$, so its lower right $(m-i) \times (m-i)$ submatrix is positive definite in this case too.

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 the execution of **modchol**. ∎

The next result bounds the remainder matrix E.

LEMMA 3.2. *Assume that* (3.10) *holds. We then have that*

$$\|E\|_2 \le \|E\|_F \le \bar{\epsilon}^{1/2}, \tag{3.11}$$

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

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

By using (3.2) and the fact that the nonzero elements of each $E^{(i)}$ occur in different locations, 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 (3.11). ∎

The bound (3.11) suggests that the matrix $\tilde{L}_{\bar{\mathcal{J}}\bar{\mathcal{J}}}^{-1} E_{\bar{\mathcal{J}}\mathcal{J}}$, which proves to be critical in our analysis, can be estimated by

$$\|\tilde{L}_{\bar{\mathcal{J}}\bar{\mathcal{J}}}^{-1} E_{\bar{\mathcal{J}}\mathcal{J}}\| \le \|\tilde{L}_{\bar{\mathcal{J}}\bar{\mathcal{J}}}^{-1}\| \|E_{\bar{\mathcal{J}}\mathcal{J}}\| \le \bar{\epsilon}^{1/2} \|\tilde{L}_{\bar{\mathcal{J}}\bar{\mathcal{J}}}^{-1}\|.$$

The following theorem shows that in fact the factor $\|\tilde{L}_{\bar{\mathcal{J}}\bar{\mathcal{J}}}^{-1}\|$ can be omitted from the right-hand side. The resulting bound is much stronger, because the omitted factor is potentially quite large.

THEOREM 3.3. *Assume that* (3.10) *holds. We then have*

$$\|\tilde{L}_{\bar{\mathcal{J}}\bar{\mathcal{J}}}^{-1} E_{\bar{\mathcal{J}}\mathcal{J}}\| \le (m\epsilon)^{1/2}. \tag{3.12}$$

Proof. We start by choosing some arbitrary index $i \in \mathcal{J}$ and examining the structure of $E_{\cdot i}$. We note from (3.1) and (3.2) that

- $E_{ji} \neq 0$ for $j < i$ only if $j \in \mathcal{J}$;
- $E_{ii} = M_{ii}^{(i-1)}$;
- $E_{ji} = M_{ji}^{(i-1)} \neq 0$ in general for all $j > i$.

Therefore, we observe that the subvector

$$E_{\bar{\mathcal{J}},i} = [E_{ji}]_{j\in\bar{\mathcal{J}}}$$

has nonzeros only in locations indexed by j with $j > i$. If we define the index subsets $\bar{\mathcal{J}}_i$ and $\mathcal{J}_i$ by

$$\bar{\mathcal{J}}_i \stackrel{\text{def}}{=} \bar{\mathcal{J}} \cap \{i+1, i+2, \dots, m\}, \qquad \mathcal{J}_i \stackrel{\text{def}}{=} \mathcal{J} \cap \{i+1, i+2, \dots, m\}, \tag{3.13}$$

it follows that

$$E_{\bar{\mathcal{J}},i} = \begin{bmatrix} 0 \\ E_{\bar{\mathcal{J}}_i,i} \end{bmatrix}. \tag{3.14}$$

It follows from this property and the lower triangularity of $\tilde{L}_{\bar{\mathcal{J}}\bar{\mathcal{J}}}$ that

$$\tilde{L}_{\bar{\mathcal{J}}\bar{\mathcal{J}}}^{-1} E_{\bar{\mathcal{J}}i} = \begin{bmatrix} 0 \\ \tilde{L}_{\bar{\mathcal{J}}_i\bar{\mathcal{J}}_i}^{-1} E_{\bar{\mathcal{J}}_i i} \end{bmatrix}. \tag{3.15}$$

From Lemma 3.1, we have that $M_{i:m,i:m}^{(i-1)}$ is symmetric positive definite. We perform symmetric permutations on this matrix to group the components in $\mathcal{J}_i$ and $\bar{\mathcal{J}}_i$, and obtain

$$\begin{bmatrix} M_{ii}^{(i-1)} & M_{i,\bar{\mathcal{J}}_i}^{(i-1)} & M_{i,\mathcal{J}_i}^{(i-1)} \\ M_{\bar{\mathcal{J}}_i,i}^{(i-1)} & M_{\bar{\mathcal{J}}_i,\bar{\mathcal{J}}_i}^{(i-1)} & M_{\bar{\mathcal{J}}_i,\mathcal{J}_i}^{(i-1)} \\ M_{\mathcal{J}_i,i}^{(i-1)} & M_{\mathcal{J}_i,\bar{\mathcal{J}}_i}^{(i-1)} & M_{\mathcal{J}_i,\mathcal{J}_i}^{(i-1)} \end{bmatrix} = \begin{bmatrix} M_{ii}^{(i-1)} & E_{\bar{\mathcal{J}}_i,i}^T & E_{\mathcal{J}_i,i}^T \\ E_{\bar{\mathcal{J}}_i,i} & M_{\bar{\mathcal{J}}_i,\bar{\mathcal{J}}_i}^{(i-1)} & M_{\bar{\mathcal{J}}_i,\mathcal{J}_i}^{(i-1)} \\ E_{\mathcal{J}_i,i} & M_{\mathcal{J}_i,\bar{\mathcal{J}}_i}^{(i-1)} & M_{\mathcal{J}_i,\mathcal{J}_i}^{(i-1)} \end{bmatrix}, \tag{3.16}$$

which is still symmetric positive definite. The principal submatrix

$$\begin{bmatrix} M_{ii}^{(i-1)} & E_{\bar{\mathcal{J}}_i,i}^T \\ E_{\bar{\mathcal{J}}_i,i} & M_{\bar{\mathcal{J}}_i,\bar{\mathcal{J}}_i}^{(i-1)} \end{bmatrix} \tag{3.17}$$

is also symmetric positive definite. It is easy to see that steps $i+1, i+2, \dots, m$ of **modchol** yield a modified Cholesky factorization of the form

$$M_{i+1:m,i+1:m}^{(i-1)} = \tilde{L}_{i+1:m,i+1:m}\tilde{L}_{i+1:m,i+1:m}^T + E_{i+1:m,i+1:m}.$$

As in (3.5a), we have that $E_{\bar{\mathcal{J}}_i,\bar{\mathcal{J}}_i} = 0$, so that by reordering and partitioning as in (3.16) and using $\tilde{L}_{\bar{\mathcal{J}}_i,\mathcal{J}_i} = 0$, we obtain

$$M_{\bar{\mathcal{J}}_i,\bar{\mathcal{J}}_i}^{(i-1)} = \tilde{L}_{\bar{\mathcal{J}}_i,\bar{\mathcal{J}}_i}\tilde{L}_{\bar{\mathcal{J}}_i,\bar{\mathcal{J}}_i}^T. \tag{3.18}$$

By the positive definite property of the matrix in (3.17), the Schur complement of $M_{ii}^{(i-1)}$ in this matrix must be positive, so we have from (3.18) that

$$0 < M_{ii}^{(i-1)} - E_{\bar{\mathcal{J}}_i,i}^T (M_{\bar{\mathcal{J}}_i,\bar{\mathcal{J}}_i}^{(i-1)})^{-1} E_{\bar{\mathcal{J}}_i,i} = M_{ii}^{(i-1)} - \|\tilde{L}_{\bar{\mathcal{J}}_i,\bar{\mathcal{J}}_i}^{-1} E_{\bar{\mathcal{J}}_i,i}\|^2.$$

Because $i \in \mathcal{J}$, we have $M_{ii}^{(i-1)} \le \epsilon$, and therefore, from (3.15), we have

$$\|\tilde{L}_{\bar{\mathcal{J}}\bar{\mathcal{J}}}^{-1} E_{\bar{\mathcal{J}},i}\| = \|\tilde{L}_{\bar{\mathcal{J}}_i,\bar{\mathcal{J}}_i}^{-1} E_{\bar{\mathcal{J}}_i,i}\| < \epsilon^{1/2}.$$

Since this bound holds for all $i \in \mathcal{J}$, we have

$$\|\tilde{L}_{\bar{\mathcal{J}}\bar{\mathcal{J}}}^{-1} E_{\bar{\mathcal{J}}\mathcal{J}}\| \le |\mathcal{J}|^{1/2}\epsilon^{1/2} \le (m\epsilon)^{1/2},$$

as required. ∎

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

THEOREM 3.4. *Suppose that* (3.10) *holds. For the exact solution z and approximate solution $\tilde{z}$ defined in* (3.8) *and* (3.9)*, respectively, we have that*

$$\|\tilde{L}^T[z-\tilde{z}]\| = \|\tilde{L}_{\bar{\mathcal{J}}\bar{\mathcal{J}}}^{-1} E_{\bar{\mathcal{J}}\mathcal{J}} z_{\mathcal{J}}\| \le (m\epsilon)^{1/2}\|z_{\mathcal{J}}\|. \tag{3.19}$$

Proof. From (3.8) together with (3.5), we have

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

while from (3.9), we have

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

By combining these two relations, we obtain

$$\tilde{L}_{\cdot\bar{\mathcal{J}}}^T[z-\tilde{z}] = -\tilde{L}_{\bar{\mathcal{J}}\bar{\mathcal{J}}}^{-1} E_{\bar{\mathcal{J}}\mathcal{J}} z_{\mathcal{J}}. \tag{3.20}$$

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

$$\sigma_1^2 \ge \sigma_2^2 \ge \cdots \ge \sigma_m^2 > 0. \tag{3.21}$$

We define the diagonal matrix Σ by

$$\Sigma = \mathrm{diag}(\sigma_1, \sigma_2, \ldots, \sigma_m). \tag{3.22}$$

It follows that there exists an orthogonal matrix Q such that

$$M = Q\Sigma^2 Q^T. \tag{3.23}$$

Because the largest diagonal in M is 1 by assumption (3.10), we have by elementary analysis that

$$1 \le \sigma_1^2 \le m. \tag{3.24}$$

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

- ϵ is somewhat smaller than 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. In particular, we show in Lemma 3.5 that they imply that $|\bar{\mathcal{J}}| \geq p$.) By partitioning the spectrum at the gap, we obtain

$$\Sigma_1 = \text{diag}(\sigma_1, \sigma_2, \ldots, \sigma_p), \qquad \Sigma_2 = \text{diag}(\sigma_{p+1}, \sigma_{p+2}, \ldots, \sigma_m). \tag{3.25}$$

From (3.23), 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

$$L^T = U \Sigma Q^T = U_1 \Sigma_1 Q_1^T + U_2 \Sigma_2 Q_2^T \tag{3.26}$$

for some $m \times m$ orthogonal matrix $U = [U_1 \,|\, U_2]$, 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 (3.4) 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

$$\tilde{\sigma}_{|\bar{\mathcal{J}}|} > \tilde{\sigma}_{|\bar{\mathcal{J}}|+1} = \cdots = \tilde{\sigma}_m = 0. \tag{3.27}$$

As in (3.26), there are orthogonal $m \times m$ matrices $\tilde{U}$ and $\tilde{Q}$ such that

$$\tilde{M} = \tilde{Q} \tilde{\Sigma}^2 \tilde{Q}^T = \tilde{Q}_1 \tilde{\Sigma}_1^2 \tilde{Q}_1^T + \tilde{Q}_2 \tilde{\Sigma}_2^2 \tilde{Q}_2^T, \tag{3.28a}$$

$$\tilde{L}^T = \tilde{U} \tilde{\Sigma} \tilde{Q}^T = \tilde{U}_1 \tilde{\Sigma}_1 \tilde{Q}_1^T + \tilde{U}_2 \tilde{\Sigma}_2 \tilde{Q}_2^T, \tag{3.28b}$$

where

$$\tilde{\Sigma}_1 = \text{diag}(\tilde{\sigma}_1, \tilde{\sigma}_2, \ldots, \tilde{\sigma}_p), \qquad \tilde{\Sigma}_2 = \text{diag}(\tilde{\sigma}_{p+1}, \ldots, \tilde{\sigma}_m), \tag{3.29}$$

with a corresponding partitioning for $\tilde{U} = [\tilde{U}_1 \,|\, \tilde{U}_2]$ and $\tilde{Q} = [\tilde{Q}_1 \,|\, \tilde{Q}_2]$. It is an immediate consequence of an eigenvalue perturbation result of Stewart and Sun [12, Corollary IV.4.13] and of our Lemma 3.2 that

$$\sum_{i=1}^{m} [\sigma_i^2 - \tilde{\sigma}_i^2]^2 \leq \|E\|_F^2 = \bar{\epsilon}. \tag{3.30}$$

The following result shows that if ϵ is sufficiently small relative to the pth eigenvalue of M, then at least p pivots are accepted during **modchol**.

LEMMA 3.5. *If $\bar{\epsilon}^{1/2} < \sigma_p^2$, we have $|\bar{\mathcal{J}}| \geq p$.*

Proof. If $|\bar{\mathcal{J}}| < p$, we have from (3.27) and (3.30) 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 $\bar{\epsilon}^{1/2} < \sigma_p^2$. ∎

Our next result concerns the differences between the subspaces spanned by Q_1 and by $\tilde{Q}_1$, the spaces of "large" eigenvalues of M and $\tilde{M}$, respectively.

LEMMA 3.6. *Suppose that* $|\bar{\mathcal{J}}| < m$ *and that the values* σ_p *and* σ_{p+1} *from* (3.21) *and* ϵ *from* **modchol** *satisfy the conditions*

$$\sigma_p^2 - \sigma_{p+1}^2 > 5\bar{\epsilon}^{1/2}. \tag{3.31}$$

Then there are matrices

$\tilde{\Lambda}_1$ $p \times p$ *symmetric positive definite,*
$\tilde{\Lambda}_2$ $(m-p) \times (m-p)$ *symmetric positive semidefinite,*
$\bar{Q}_1$ $m \times p$ *orthonormal,*
$\bar{Q}_2$ $m \times (m-p)$ *orthonormal,*

such that

$$\tilde{M} = \bar{Q}\tilde{\Lambda}\bar{Q}^T = \bar{Q}_1\tilde{\Lambda}_1\bar{Q}_1^T + \bar{Q}_2\tilde{\Lambda}_2\bar{Q}_2^T, \tag{3.32}$$

$$\|\bar{Q}_1 - Q_1\| \le \frac{2\bar{\epsilon}^{1/2}}{\sigma_p^2 - \sigma_{p+1}^2 - 2\bar{\epsilon}^{1/2}}, \tag{3.33}$$

$$\|\tilde{\Lambda}_1 - \Sigma_1^2\| \le 2\bar{\epsilon}^{1/2}, \tag{3.34}$$

$$\|\tilde{\Lambda}_2 - \Sigma_2^2\| \le 2\bar{\epsilon}^{1/2}, \tag{3.35}$$

where

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

Moreover, there are matrices

$V_1 p \times p$ *orthogonal,*
$V_2 (m-p) \times (m-p)$ *orthogonal,*

such that

$$\tilde{\Sigma}_1^2 = V_1^T\tilde{\Lambda}_1V_1, \tilde{Q}_1 = \bar{Q}_1V_1, \tag{3.36a}$$

$$\tilde{\Sigma}_2^2 = V_2^T\tilde{\Lambda}_2V_2, \tilde{Q}_2 = \bar{Q}_2V_2, \tag{3.36b}$$

where $\tilde{\Sigma}$ *and* $\tilde{Q}$ *are defined as in* (3.28).

Proof. Note first that $p \le |\bar{\mathcal{J}}|$ by (3.31) and Lemma 3.5. The result is a straightforward consequence of Theorem V.2.8 of Stewart and Sun [12, p. 238]. Since $\tilde{M} = M - E$, we use (3.23) and partition as in (3.25) to obtain

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

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

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

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

$$\bar{Q}_1 = Q_1 + Q_2P \tag{3.37}$$

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

$$\|P\| \le \frac{\tilde{\gamma}}{\tilde{\delta}} \le \frac{2\bar{\epsilon}^{1/2}}{\sigma_p^2 - \sigma_{p+1}^2 - 2\bar{\epsilon}^{1/2}} < 1. \tag{3.38}$$

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

$$\bar{Q}_1^T \tilde{M} \bar{Q}_1 = \tilde{\Lambda}_1 = \Sigma_1^2 + F_{11} + F_{12}P. \tag{3.39}$$

The bound (3.33) follows from (3.37), (3.38), and $\|Q_2\| = 1$. It follows immediately from the first equality in (3.39) that $\tilde{\Lambda}_1$ is symmetric, and we have

$$\|\tilde{\Lambda}_1 - \Sigma_1^2\| \le \|F_{11}\| + \|F_{12}\|\|P\| \le 2\bar{\epsilon}^{1/2}, \tag{3.40}$$

verifying the inequality (3.34). This inequality implies that the smallest singular value of $\tilde{\Lambda}$ is no smaller than $\sigma_p^2 - 2\bar{\epsilon}^{1/2}$, which by (3.31) is positive, so $\tilde{\Lambda}_1$ is symmetric positive definite.

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

$$\bar{Q}_2^T \tilde{M} \bar{Q}_2 = \tilde{\Lambda}_2.$$

Symmetric positive semidefiniteness of $\tilde{\Lambda}_2$ follows immediately. By using the invariant subspace property, we obtain

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

from which (3.32) follows immediately.

Similarly to (3.40), we have that

$$\|\tilde{\Lambda}_2 - \Sigma_2^2\| \le 2\bar{\epsilon}^{1/2},$$

so the largest eigenvalue of $\tilde{\Lambda}_2$ is no larger than $\sigma_{p+1}^2 + 2\bar{\epsilon}^{1/2}$. Because of (3.31) and our earlier observation that the smallest eigenvalue of $\tilde{\Lambda}_1$ is no smaller than $\sigma_p^2 - 2\bar{\epsilon}^{1/2}$, we conclude that the eigenvalues of $\tilde{\Lambda}_1$ are the p largest eigenvalues $\tilde{\sigma}_1^2, \tilde{\sigma}_2^2, \ldots, \tilde{\sigma}_p^2$, while those of $\tilde{\Lambda}_2$ are the $(m-p)$ smallest eigenvalues. By our definition (3.29), we conclude that there are orthogonal matrices V_1 and V_2 such that

$$V_1 \tilde{\Sigma}_1^2 V_1^T = \tilde{\Lambda}_1 \quad \text{and} \quad V_2 \tilde{\Sigma}_2^2 V_2^T = \tilde{\Lambda}_2.$$

By substituting these expressions into (3.32) and setting $\tilde{Q}_1 = \bar{Q}_1 V_1$ and $\tilde{Q}_2 = \bar{Q}_2 V_2$, we recover (3.28a). $\square$

Lemma 3.6 suggests a few other estimates and assumptions that will be useful in subsequent sections. When (3.31) holds, we have from (3.30) that

$$\tilde{\sigma}_1^2 \le \sigma_1^2 + \bar{\epsilon}^{1/2} < \sigma_1^2 + .2\sigma_p^2 < 1.2\sigma_1^2 \le 1.2m \tag{3.41}$$

(where the last inequality follows from (3.24)), and also that

$$\tilde{\sigma}_p^2 \ge \sigma_p^2 - \bar{\epsilon}^{1/2} \ge .8\sigma_p^2 \;\;\Rightarrow\;\; \tilde{\sigma}_p^{-1} \le 1.2\sigma_p^{-1}. \tag{3.42}$$

When we make the additional assumption that

$$\frac{\sigma_{p+1}^2}{\sigma_p^2} \le \frac{1}{10} \tag{3.43}$$

(indicating that the gap in the spectrum actually separates the small and large eigenvalues), we derive that

$$\begin{aligned} \|\bar{Q}_1 - Q_1\| &\le \frac{2\bar{\epsilon}^{1/2}}{\sigma_p^2 - \sigma_{p+1}^2 - 2\bar{\epsilon}^{1/2}} \\ &= \frac{2\bar{\epsilon}^{1/2}}{\sigma_p^2}\left[1 - \frac{\sigma_{p+1}^2}{\sigma_p^2} - 2\frac{\bar{\epsilon}^{1/2}}{\sigma_p^2}\right]^{-1} \\ &\le \frac{2\bar{\epsilon}^{1/2}}{\sigma_p^2}[1 - .1 - .4]^{-1} \le \frac{4\bar{\epsilon}^{1/2}}{\sigma_p^2}. \end{aligned} \tag{3.44}$$

Another useful quantity that enters into our error bounds is the norm of $\tilde{L}_{\bar{\mathcal{J}}\bar{\mathcal{J}}}^{-1}$, which we denote by τ; that is,

$$\tau \stackrel{\text{def}}{=} \|\tilde{L}_{\bar{\mathcal{J}}\bar{\mathcal{J}}}^{-1}\| = \sigma_{|\bar{\mathcal{J}}|}(\tilde{L}_{\bar{\mathcal{J}}\bar{\mathcal{J}}})^{-1}, \tag{3.45}$$

where $\sigma_{|\bar{\mathcal{J}}|}(\tilde{L}_{\bar{\mathcal{J}}\bar{\mathcal{J}}})$ denotes the $|\bar{\mathcal{J}}|$th singular value of $\tilde{L}_{\bar{\mathcal{J}}\bar{\mathcal{J}}}$. Because of (3.5a) and the fact that all diagonals of $M_{\bar{\mathcal{J}}\bar{\mathcal{J}}}$ are bounded by 1 (by our assumption (3.10)), we have that $\sigma_{|\bar{\mathcal{J}}|}(\tilde{L}_{\bar{\mathcal{J}}\bar{\mathcal{J}}}) \le 1$ and therefore that

$$\tau \ge 1. \tag{3.46}$$

Using (3.5a) again, we have that

$$\|M_{\bar{\mathcal{J}}\bar{\mathcal{J}}}^{-1}\| = \|\tilde{L}_{\bar{\mathcal{J}}\bar{\mathcal{J}}}^{-1}\|^2 = \tau^2. \tag{3.47}$$

Since $\|M_{\bar{\mathcal{J}}\bar{\mathcal{J}}}\| \le \|M\| \le \sigma_1^2$, we have from (3.24) and (3.47) that

$$\kappa(M_{\bar{\mathcal{J}}\bar{\mathcal{J}}}) \le \sigma_1^2\tau^2 \le m\tau^2. \tag{3.48}$$

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 (3.9), where the subvector $\tilde{z}_{\bar{\mathcal{J}}}$ satisfies the system

$$M_{\bar{\mathcal{J}}\bar{\mathcal{J}}}\tilde{z}_{\bar{\mathcal{J}}} = \tilde{L}_{\bar{\mathcal{J}}\bar{\mathcal{J}}}\tilde{L}_{\bar{\mathcal{J}}\bar{\mathcal{J}}}^T\tilde{z}_{\bar{\mathcal{J}}} = r_{\bar{\mathcal{J}}}, \tag{4.1}$$

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 **modchol**,
- error arising during the triangular substitutions in (4.1), and
- error in the evaluation of the matrix M and the right-hand-side r.

Since **modchol** amounts to a standard Cholesky factorization/triangular-solve procedure on the matrix $M_{\bar{\mathcal{J}}\bar{\mathcal{J}}}$, roundoff error in **modchol** and errors arising during the triangular substitutions can all be accounted for by adding a term $E^{\mathbf{u}}_{\bar{\mathcal{J}}\bar{\mathcal{J}}}$ to the coefficient matrix $M_{\bar{\mathcal{J}}\bar{\mathcal{J}}}$ in (4.1), where

$$\|E^{\mathbf{u}}_{\bar{\mathcal{J}}\bar{\mathcal{J}}}\| \le \delta_{\mathbf{u}} \|M_{\bar{\mathcal{J}}\bar{\mathcal{J}}}\| \le \delta_{\mathbf{u}}; \tag{4.2}$$

see, for example, Higham [7, Theorem 10.4]. (Recall from section 1 that $\delta_{\mathbf{u}}$ denotes a modest multiple of $\mathbf{u}$ and that $\|M_{\bar{\mathcal{J}}\bar{\mathcal{J}}}\| \le \sqrt{n}$ because of (3.10).) We assume that the error in evaluating M can also be incorporated into $E^{\mathbf{u}}_{\bar{\mathcal{J}}\bar{\mathcal{J}}}$; this is certainly true in section 5, for instance. As we see in this section, the remaining source of error—the error that arises in evaluation of the right-hand side—plays a significant role in the interior-point application. Our results are strengthened if we account for some of this error by placing it explicitly in the range space of L; that is, we write it as $Lf + e$, for some vectors f and e. (We refer to e as the "unstructured error.") The computed solution $\hat{z}_{\bar{\mathcal{J}}}$ of the system (4.1) therefore satisfies

$$(M_{\bar{\mathcal{J}}\bar{\mathcal{J}}} + E^{\mathbf{u}}_{\bar{\mathcal{J}}\bar{\mathcal{J}}})\hat{z}_{\bar{\mathcal{J}}} = (r + Lf + e)_{\bar{\mathcal{J}}}. \tag{4.3}$$

The following result shows that we can repartition the right-hand-side error according to the approximate Cholesky factor $\tilde{L}$, a fact that is useful in the main error results of this section.

LEMMA 4.1. *Suppose that* (3.10), (3.31), *and* (3.43) *hold. Given vectors* $e, f \in \mathsf{R}^m$, *we have*

$$Lf + e = \tilde{L}\tilde{f} + \tilde{e}, \tag{4.4}$$

where

$$\|\tilde{f}\| \le \delta_1 \sigma_p^{-1} \|f\|, \qquad \|\tilde{e}\| \le \delta_1 \left(\bar{\epsilon}^{1/2}\sigma_p^{-3} + \sigma_{p+1}\right)\|f\| + \|e\|. \tag{4.5}$$

Proof. From (3.26), we have

$$Lf + e = Q_1\Sigma_1 U_1^T f + Q_2\Sigma_2 U_2^T f + e = Q_1\Sigma_1^2 f_1 + e_1,$$

where the vectors f_1 and e_1 defined by

$$f_1 = \Sigma_1^{-1}U_1^T f, \qquad e_1 = Q_2\Sigma_2 U_2^T f + e$$

satisfy the bounds

$$\|f_1\| \le \sigma_p^{-1}\|f\|, \qquad \|e_1\| \le \sigma_{p+1}\|f\| + \|e\|; \tag{4.6}$$

see (3.25). Using the notation of (3.28), (3.29), and (3.32), we define the vector $\tilde{e}$ by

$$\tilde{e} = (Q_1 - \bar{Q}_1)\tilde{\Lambda}_1 f_1 + Q_1(\Sigma_1^2 - \tilde{\Lambda}_1)f_1 + e_1$$

and note that

$$Lf + e = Q_1\Sigma_1^2 f_1 + e_1 = \bar{Q}_1\tilde{\Lambda}_1 f_1 + \tilde{e}. \tag{4.7}$$

By using (3.34), (3.41), (3.44), and (4.6), we can bound the terms of $\tilde{e}$ to obtain

$$\begin{aligned}\|\tilde{e}\| &\le \|Q_1 - \bar{Q}_1\|\,\|\tilde{\Lambda}_1\|\,\|f_1\| + \|\Sigma_1^2 - \tilde{\Lambda}_1\|\|f_1\| + \|e_1\| \\ &\le 4\frac{\bar{\epsilon}^{1/2}}{\sigma_p^2}(1.2\sigma_1^2)\sigma_p^{-1}\|f\| + 2\bar{\epsilon}^{1/2}\sigma_p^{-1}\|f\| + \sigma_{p+1}\|f\| + \|e\|,\end{aligned}$$

from which the bound in (4.5) follows if we use the inequality (3.24). For the companion term on the right-hand side of (4.7), we have from (3.36) that

$$\bar{Q}_1\tilde{\Lambda}_1 f_1 = \bar{Q}_1 V_1(V_1^T\tilde{\Lambda}_1 V_1)(V_1^T f_1) = \tilde{Q}_1\tilde{\Sigma}_1(\tilde{\Sigma}_1 V_1^T f_1).$$

Using $\tilde{U}$ defined in (3.28b), we set

$$\tilde{f} = [\tilde{U}_1 \,|\, \tilde{U}_2]\begin{bmatrix} \tilde{\Sigma}_1 V_1^T f_1 \\ 0 \end{bmatrix},$$

so from (3.28b) and (3.36a), we obtain that

$$\tilde{L}\tilde{f} = \tilde{Q}_1\tilde{\Sigma}_1\tilde{U}_1^T\tilde{f} + \tilde{Q}_2\tilde{\Sigma}_2\tilde{U}_2^T\tilde{f} = \tilde{Q}_1\tilde{\Sigma}_1(\tilde{\Sigma}_1 V_1^T f_1) = \bar{Q}_1\tilde{\Lambda}_1 f_1.$$

Hence, by substituting in (4.7), we obtain $Lf+e = \tilde{L}\tilde{f}+\tilde{e}$. To obtain the bound on $\|\tilde{f}\|$, we simply use its definition above together with (3.41), (4.6), and the orthonormality of $\tilde{U}_1$ and V_1. $\square$

Before stating our main result, we introduce two additional assumptions. The first is that finite precision does not affect cutoff decisions in **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. Provided that we have

$$\epsilon \geq 100\mathbf{u}, \tag{4.8}$$

say, the role of this assumption is to provide a convenient link between the results of sections 3 and 4. It is not really essential to the analysis, for reasons that we now explain. We can show by a standard error analysis argument that the matrix $\tilde{L}$ obtained in finite-precision arithmetic is the same as the one we would obtain by applying **modchol** in exact arithmetic to a perturbed matrix $M + \hat{E}^{\mathbf{u}}$, where $\|\hat{E}^{\mathbf{u}}\| \leq \delta_{\mathbf{u}}\|M\| \leq \delta_{\mathbf{u}}$. Hence, finite-precision arithmetic causes changes of size $\delta_{\mathbf{u}}$ in the diagonal elements that are tested against the threshold $\beta\epsilon$ in **modchol**. If $\mathbf{u}$ is significantly less than $\beta\epsilon$ (as in (4.8)), only a few skipping decisions would be affected by this perturbation. Moreover, we could generalize the analysis of section 3 so that it applies to the slightly perturbed matrix $M + \hat{E}^{\mathbf{u}}$ rather than to the exact matrix M, hence ensuring that the results of that section apply to the set $\mathcal{J}$ calculated in a finite-precision environment. We prefer to avoid the additional complication, however, and simply assume that the sets $\mathcal{J}$ that we discuss in sections 3 and 4 are one and the same. In any case, we note that when $\bar{\epsilon}$ falls in the gap between large and small eigenvalues, the makeup of $\mathcal{J}$ is not affected at all.

The second assumption is that

$$\tau\bar{\epsilon}^{1/2} = \delta_1. \tag{4.9}$$

We can expect this estimate to hold in all but pathological cases, since the elements of $\tilde{L}_{\bar{\mathcal{J}}\bar{\mathcal{J}}}$ are bounded by 1, and its diagonal elements lie in the range $[\bar{\epsilon}^{1/2}, 1]$.

In the following result, we bound the difference $L^T(\hat{z}-z)$ in terms of $\|\hat{z}\|$, $\|z\|$, and the norms $\|f\|$ and $\|e\|$ of the perturbation vectors. The explicit appearance of the computed solution $\|\hat{z}\|$ in the right-hand-side bound is not standard practice in error analysis, but we were motivated to include it by our numerical experience on practical linear programming problems. We can derive a rigorous bound on $\|\hat{z}\|$ in terms of $\|z\|$, $\|f\|$, and $\|e\|$, but numerical experience shows that this bound appears to be too pessimistic, so it turns out to be more illuminating to leave $\|\hat{z}\|$ in place and to work with a direct estimate of this quantity.

THEOREM 4.2. *Suppose that $\hat{z}_{\bar{\mathcal{J}}}$ solves (4.3), where $E^{\mathbf{u}}_{\bar{\mathcal{J}}\bar{\mathcal{J}}}$ is bounded as in (4.2). Suppose too that we set $\hat{z}_{\mathcal{J}} = 0$ (as in (3.9)), that (3.10), (3.31), (4.9), and (3.43) hold, and that roundoff error does not affect the composition of $\mathcal{J}$. We then have*

$$\|L^T(\hat{z}-z)\| \le \delta_1 \left[\sigma_p^{-2}(\tau\mathbf{u} + \bar{\epsilon}^{1/2}) + \sigma_{p+1}\right] \|\hat{z}\| + \delta_1 \left[\sigma_p^{-2}\bar{\epsilon}^{1/2} + \sigma_{p+1}\right] \|z\| \\ + \delta_1 \left(\sigma_p^{-4} + \tau\sigma_{p+1}\sigma_p^{-1}\right) \|f\| + \delta_1 \tau \sigma_p^{-1} \|e\|, \tag{4.10}$$

where z is the exact solution from (3.7). In the special case of $J = \emptyset$, we have

$$\|L^T(\hat{z}-z)\| \le \tau\delta_{\mathbf{u}}\sigma_1\|\hat{z}\| + \|f\| + \tau\|e\|. \tag{4.11}$$

Proof. From (3.26), we have

$$\begin{aligned} \|L^T(\hat{z}-z)\| &= \left\| \begin{bmatrix} \Sigma_1 Q_1^T(\hat{z}-z) \\ \Sigma_2 Q_2^T(\hat{z}-z) \end{bmatrix} \right\| \\ &\le \|\Sigma_1\|\, \|Q_1^T(\hat{z}-z)\| + \|\Sigma_2\|\, \|\hat{z}-z\| \\ &\le \|\Sigma_1\|\, \|\bar{Q}_1^T(\hat{z}-z)\| + \|\Sigma_1\|\, \|Q_1 - \bar{Q}_1\| \|\hat{z}-z\| + \|\Sigma_2\|\, \|\hat{z}-z\|. \end{aligned} \tag{4.12}$$

To bound the first term, we note from (3.28b) that

$$\|\tilde{L}^T(\hat{z}-z)\| = \left\| \begin{bmatrix} \tilde{\Sigma}_1 \tilde{Q}_1^T(\hat{z}-z) \\ \tilde{\Sigma}_2 \tilde{Q}_2^T(\hat{z}-z) \end{bmatrix} \right\|,$$

and therefore, from (3.36a) and (3.29), we have

$$\|\bar{Q}_1^T(\hat{z}-z)\| = \|\tilde{Q}_1^T(\hat{z}-z)\| \le \|\tilde{\Sigma}_1^{-1}\|\, \|\tilde{\Sigma}_1\tilde{Q}_1^T(\hat{z}-z)\| \le \tilde{\sigma}_p^{-1}\|\tilde{L}^T(\hat{z}-z)\|. \tag{4.13}$$

Since $\tilde{L}_{\cdot\mathcal{J}} = 0$ and $\hat{z}_{\mathcal{J}} = 0$, we have too that

$$\tilde{L}^T(\hat{z}-z) = \tilde{L}^T_{\bar{\mathcal{J}}\bar{\mathcal{J}}}(\hat{z}_{\bar{\mathcal{J}}} - z_{\bar{\mathcal{J}}}) - \tilde{L}^T_{\mathcal{J}\bar{\mathcal{J}}} z_{\mathcal{J}}. \tag{4.14}$$

By substituting (3.42) and (4.14) into (4.13), we obtain

$$\|\bar{Q}_1^T(\hat{z}-z)\| \le 1.2\sigma_p^{-1}\|\tilde{L}^T_{\bar{\mathcal{J}}\bar{\mathcal{J}}}(\hat{z}_{\bar{\mathcal{J}}} - z_{\bar{\mathcal{J}}}) - \tilde{L}^T_{\mathcal{J}\bar{\mathcal{J}}} z_{\mathcal{J}}\|. \tag{4.15}$$

From (4.3) and (4.4), and using (3.5a) and $\tilde{L}_{\cdot\mathcal{J}} = 0$, we have that

$$(\tilde{L}_{\bar{\mathcal{J}}\bar{\mathcal{J}}}\tilde{L}^T_{\bar{\mathcal{J}}\bar{\mathcal{J}}} + E^{\mathbf{u}}_{\bar{\mathcal{J}}\bar{\mathcal{J}}})\hat{z}_{\bar{\mathcal{J}}} = r_{\bar{\mathcal{J}}} + \tilde{L}_{\bar{\mathcal{J}}\bar{\mathcal{J}}}\tilde{f}_{\bar{\mathcal{J}}} + \tilde{e}_{\bar{\mathcal{J}}}.$$

Meanwhile, from (3.5) and (3.7), we have

$$\tilde{L}_{\bar{\mathcal{J}}\bar{\mathcal{J}}}\tilde{L}^T_{\bar{\mathcal{J}}\bar{\mathcal{J}}} z_{\bar{\mathcal{J}}} + \tilde{L}_{\bar{\mathcal{J}}\bar{\mathcal{J}}}\tilde{L}^T_{\mathcal{J}\bar{\mathcal{J}}} z_{\mathcal{J}} + E_{\bar{\mathcal{J}}\mathcal{J}} z_{\mathcal{J}} = r_{\bar{\mathcal{J}}}.$$

By combining these two equations and multiplying by $\tilde{L}^{-1}_{\bar{\mathcal{J}}\bar{\mathcal{J}}}$, we obtain

$$\tilde{L}^T_{\bar{\mathcal{J}}\bar{\mathcal{J}}}(\hat{z}_{\bar{\mathcal{J}}} - z_{\bar{\mathcal{J}}}) - \tilde{L}^T_{\mathcal{J}\bar{\mathcal{J}}} z_{\mathcal{J}} = -\tilde{L}^{-1}_{\bar{\mathcal{J}}\bar{\mathcal{J}}} E^{\mathbf{u}}_{\bar{\mathcal{J}}\bar{\mathcal{J}}}\hat{z}_{\bar{\mathcal{J}}} + \tilde{L}^{-1}_{\bar{\mathcal{J}}\bar{\mathcal{J}}} E_{\bar{\mathcal{J}}\mathcal{J}} z_{\mathcal{J}} + \tilde{f}_{\bar{\mathcal{J}}} + \tilde{L}^{-1}_{\bar{\mathcal{J}}\bar{\mathcal{J}}}\tilde{e}_{\bar{\mathcal{J}}}.$$

By substituting into (4.15), and using the bounds (3.45), (3.12), and (4.2), we obtain

$$\|\bar{Q}_1^T(\hat{z}-z)\| \le \tau\delta_{\mathbf{u}}\sigma_p^{-1}\|\hat{z}_{\bar{\mathcal{J}}}\| + \delta_1\bar{\epsilon}^{1/2}\|z_{\mathcal{J}}\| + \|\tilde{f}_{\bar{\mathcal{J}}}\| + \tau\|\tilde{e}_{\bar{\mathcal{J}}}\|. \tag{4.16}$$

Turning now to the second and third terms in (4.12), we have from (3.25) that

$$\|\Sigma_1\| = \sigma_1 = \delta_1, \qquad \|\Sigma_2\| = \sigma_{p+1}. \tag{4.17}$$

By substituting (4.15), (4.16), (4.17), and (3.44) into (4.12), and using

$$\|\hat{z} - z\| \leq \|\hat{z}\| + \|z\|, \qquad \|\hat{z}_{\bar{\mathcal{J}}}\| \leq \|\hat{z}\|, \qquad \|z_{\mathcal{J}}\| \leq \|z\|, \qquad 1 \leq \delta_1 \sigma_p^{-1},$$

we obtain

$$\begin{aligned}
&\|L^T(\hat{z} - z)\| \\
&\leq \delta_1 \sigma_p^{-1} \left[\tau \mathbf{u} \|\hat{z}\| + \bar{\epsilon}^{1/2} \|z\| + \|\tilde{f}\| + \tau \|\tilde{e}\| \right] + \delta_1 \left(\sigma_p^{-2} \bar{\epsilon}^{1/2} + \sigma_{p+1} \right) (\|\hat{z}\| + \|z\|) \\
&\leq \delta_1 \left[\sigma_p^{-2} (\tau \mathbf{u} + \bar{\epsilon}^{1/2}) + \sigma_{p+1} \right] \|\hat{z}\| + \delta_1 \left[\sigma_p^{-2} \bar{\epsilon}^{1/2} + \sigma_{p+1} \right] \|z\| \\
&\quad + \delta_1 \sigma_p^{-1} \|\tilde{f}\| + \delta_1 \tau \sigma_p^{-1} \|\tilde{e}\|.
\end{aligned} \tag{4.18}$$

By substituting from (4.5) and using (4.9), we have

$$\delta_1 \sigma_p^{-1} \|\tilde{f}\| + \delta_1 \tau \sigma_p^{-1} \|\tilde{e}\| \leq \delta_1 \left(\sigma_p^{-4} + \tau \sigma_{p+1} \sigma_p^{-1} \right) \|f\| + \delta_1 \tau {\sigma_p}^{-1} \|e\|.$$

By substituting into (4.18), we obtain (4.10).

For the case of $\mathcal{J} = \emptyset$, we have

$$\tilde{L}_{\bar{\mathcal{J}}\bar{\mathcal{J}}} = \tilde{L} = L, \qquad \hat{z}_{\bar{\mathcal{J}}} = \hat{z}, \qquad z_{\bar{\mathcal{J}}} = z, \qquad z_{\mathcal{J}} \ \text{vacuous},$$

while from (4.4), we have $\tilde{f} = f$, $\tilde{e} = e$. By using these equivalences in (4.16), we obtain the result (4.11) directly. ∎

Note that in the case of $\mathcal{J} = \emptyset$, we have from (3.45) that

$$\tau = \|L^{-1}\| = \sigma_m^{-1},$$

so it follows from (4.11) that

$$\|\hat{z} - z\| \leq \sigma_m^{-2} \delta_{\mathbf{u}} \|\hat{z}\| + \sigma_m^{-1} \|f\| + \sigma_m^{-2} \|e\|.$$

If we put all the right-hand-side perturbations into the vector e, and set $f = 0$, we can use the relation $\|M^{-1}\| = \sigma_m^{-2}$ to obtain

$$\|\hat{z} - z\| \leq \|M^{-1}\| \left(\delta_{\mathbf{u}} \|\hat{z}\| + \|e\| \right),$$

which is a perturbation bound for (4.3) of the type that is usually found in the numerical analysis literature.

5. Application to the interior-point algorithm. We now return to the motivating application: primal-dual interior-point algorithms for linear programming and, in particular, the linear system (2.15) 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 method, during which μ is small and the ill conditioning of AD^2A^T can become acute. Our results show where errors arise in $(\widehat{\Delta x}, \widehat{\Delta \pi}, \widehat{\Delta s})$, what effect these errors have on the steplength and the computed residual vectors r_b and r_c, and the accuracy that can be attained by the interior-point algorithm in finite precision. They also suggest a choice for the parameter ϵ and for the termination criterion.

Throughout this section, we use an informal style of analysis that combines the use of δ_1 and order notation defined in section 1. Specifically, we often replace the estimate $v = O(\epsilon)$ by $v = \delta_1\epsilon$ instead. This convention allows us to analyze the dependence of certain quantities on the unit roundoff $\mathbf{u}$ and the duality measure μ jointly.

5.1. Size estimate for a general step. We start by estimating the sizes of the various constituents of the equations (2.15)—the residuals r_b and r_c of (2.7), the $\mathcal{B}$ and $\mathcal{N}$ components of x, s, and the diagonal matrix D. Each iterate (x, π, s) of a typical primal-dual interior-point iterate satisfies the following estimates (see, for example, S. J. Wright [17]):

$$\begin{aligned} \|r_b\| &= O(\mu), & \|r_c\| &= O(\mu), \\ x_i &= \Theta(1)\ \ (i \in \mathcal{B}), & x_i &= \Theta(\mu)\ \ (i \in \mathcal{N}), \\ s_i &= \Theta(\mu)\ \ (i \in \mathcal{B}), & s_i &= \Theta(1)\ \ (i \in \mathcal{N}). \end{aligned} \tag{5.1}$$

In theoretical algorithms, these estimates follow from a requirement that all iterates must belong to a certain neighborhood of the central trajectory. In practical algorithms, the conditions for membership of the neighborhood are rarely checked explicitly, but the estimates (5.1) are still observed to hold on the vast majority of practical problems in which the primal-dual solution set is nonempty and bounded. An immediate consequence of these estimates and the definition (2.14) is that

$$D_{ii}^2 = \Theta(\mu^{-1})\ \ (i \in \mathcal{B}), \qquad D_{ii}^2 = \Theta(\mu)\ \ (i \in \mathcal{N}). \tag{5.2}$$

As mentioned in section 2, we assume that A has full rank.

We analyze a general step $(\Delta x, \Delta\pi, \Delta s)$ that satisfies the system (2.6), where r_b and r_c are given by (2.7) while r_{xs} has the form

$$r_{xs} = XS\mathbf{1} + w \quad \text{for some } w \text{ satisfying } w = O(\mu^2). \tag{5.3}$$

It is not difficult to show that the resulting step satisfies the estimate

$$(\Delta x, \Delta\pi, \Delta s) = O(\mu) \tag{5.4}$$

by using an argument based on splitting the step into an affine-scaling component $(\Delta x^{\text{aff}}, \Delta\pi^{\text{aff}}, \Delta s^{\text{aff}})$ of the step (obtained by setting $w = 0$; see (2.8)) and a "remainder" component $(\Delta x^w, \Delta\pi^w, \Delta s^w)$ that satisfies

$$\begin{bmatrix} 0 & A^T & I \\ A & 0 & 0 \\ S & 0 & X \end{bmatrix} \begin{bmatrix} \Delta x^w \\ \Delta\pi^w \\ \Delta s^w \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \\ -w \end{bmatrix}. \tag{5.5}$$

We have from [17, Theorem 7.5] that

$$\|(\Delta x^{\text{aff}}, \Delta s^{\text{aff}})\| = O(\mu), \tag{5.6}$$

while from (2.15b) and (5.1), we have

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

and since A has full rank, we have $\Delta\pi^{\text{aff}} = O(\mu)$ as well. By performing block elimination on (5.5), we have that

$$AD^2A^T\Delta\pi^w = AD^2(X^{-1}w).$$

A well-known result (see Stewart [11], Todd [13], Dikin [4], and Vanderbei and Lagarias [14]) states that the norm $\|(AD^2A^T)^{-1}AD^2\|$ is bounded over the set of all positive definite diagonal matrices D. Therefore, we have that

$$\|\Delta\pi^w\| = O(\|X^{-1}w\|).$$

From (5.1), we have $\|X^{-1}\| = O(\mu^{-1})$, so from $w = O(\mu^2)$ it follows that $\Delta\pi^w = O(\mu)$. Similar arguments based on the Stewart–Todd result can be used to show that

$$\|\Delta x^w\| = O(\mu), \qquad \|\Delta s^w\| = O(\mu).$$

The general choice (5.3) of w encompasses the affine-scaling method (2.8), for which $w = 0$. It also includes as a special case the path-following choice (2.9) when $\zeta = O(\mu)$, which can be assumed to hold on the late iterations of a superlinearly convergent method. Finally, it usually includes the Mehrotra method (2.11), since by (5.6) we have that $\|\Delta X^{\text{aff}}\Delta S^{\text{aff}}\mathbf{1}\| = O(\mu^2)$, while the heuristic choice of the parameter ζ is usually chosen by a heuristic that ensures that $\zeta = O(\mu)$.

5.2. Steplength along the exact step. We have noted already in (5.4) that $(\Delta x, \Delta\pi, \Delta s) = O(\mu)$. We can be more specific about the sizes of the critical components Δx_i, $i \in \mathcal{N}$, and Δs_i, $i \in \mathcal{B}$. If we multiply the third block row in (2.6) by $(XS)^{-1}$, use the definition (5.3), and note from (5.1) that $(x_is_i)^{-1} = \Theta(\mu^{-1})$ for $i = 1, 2, \ldots, n$, we obtain

$$\frac{\Delta x_i}{x_i} + \frac{\Delta s_i}{s_i} = -1 + O(\mu), \qquad i = 1, 2, \ldots, n.$$

Therefore, from (5.1) and (5.4), we have for $i \in \mathcal{N}$ that

$$\frac{\Delta x_i}{x_i} = -1 + \frac{O(\mu)}{\Theta(1)} = -1 + O(\mu),$$

and therefore, using (5.1) again, we have

$$\Delta x_i = -x_i + O(\mu^2), \qquad i \in \mathcal{N}. \tag{5.7}$$

In a similar way, we obtain

$$\Delta s_i = -s_i + O(\mu^2), \qquad i \in \mathcal{B}. \tag{5.8}$$

From the estimates (5.4), (5.7), and (5.8), we can show that a near-unit step can be taken along the direction $(\Delta x, \Delta\pi, \Delta s)$ without violating positivity of the x and s components. By substituting in (2.13), we can show that

$$1 - \alpha_{\max} = O(\mu). \tag{5.9}$$

To verify this estimate, suppose that $s_i + \alpha\Delta s_i = 0$ for some index $i \in \mathcal{B}$. From (5.8), we have

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

so it follows from (5.1) that

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

For the corresponding component x_i, we have from (5.1) and (5.4) that $x_i = \Theta(1)$ and $\Delta x_i = O(\mu)$. Hence, for all μ sufficiently small and all $\alpha \in [0,1]$, we have $x_i + \alpha \Delta x_i > 0$. Similar logic can be applied to the remaining indices $i \in \mathcal{N}$, thereby proving (5.9).

5.3. Scaling the system (2.15a). We can use (5.2) to analyze the eigenstructure of the coefficient matrix AD^2A^T. We have

$$AD^2A^T = A_{\cdot\mathcal{B}} D_{\mathcal{B}}^2 A_{\cdot\mathcal{B}}^T + A_{\cdot\mathcal{N}} D_{\mathcal{N}}^2 A_{\cdot\mathcal{N}}^T,$$

where the first term on the right-hand side is a matrix whose rank is $\operatorname{rank} A_{\cdot\mathcal{B}}$ in which all the nonzero eigenvalues are of size $\Theta(\mu^{-1})$. By combining this observation with the full-rank assumption on A, we obtain that

$$\sigma_i(AD^2A^T) = \Theta(\mu^{-1}), \qquad i = 1, 2, \ldots, \operatorname{rank} A_{\cdot\mathcal{B}}, \tag{5.10a}$$

$$\sigma_i(AD^2A^T) = \Theta(\mu), \qquad i = \operatorname{rank} A_{\cdot\mathcal{B}} + 1, \ldots, m. \tag{5.10b}$$

To ensure (3.10), we work with a scaled version of the matrix AD^2A^T, in which the scaling factor ρ is chosen as

$$\rho = \left[\max_{i=1,2,\ldots,m} (AD^2A^T)_{ii} \right]^{-1}. \tag{5.11}$$

Obviously, we have $\rho = \Theta(\mu)$, and by choosing p (see section 3) as

$$p = \operatorname{rank} A_{\cdot\mathcal{B}}, \tag{5.12}$$

we find that the eigenvalues $\sigma_1^2, \sigma_2^2, \ldots, \sigma_m^2$ of ρAD^2A^T satisfy

$$\sigma_i^2 = \Theta(1), \qquad i = 1, 2, \ldots, p, \tag{5.13a}$$

$$\sigma_i^2 = \Theta(\mu^2), \qquad i = p+1, \ldots, m. \tag{5.13b}$$

The exact Cholesky factor L satisfies

$$LL^T = \rho AD^2A^T. \tag{5.14}$$

Suppose now that **modchol** is used to compute the solution of the scaled version of the system (2.15a), namely,

$$\rho AD^2A^T \Delta\pi = -\rho r_b - \rho AD^2(r_c - X^{-1} r_{xs}), \tag{5.15}$$

where r_{xs} is defined as in (5.3). This process is carried out in finite-precision arithmetic, resulting in a computed solution $\widehat{\Delta\pi}$. The remaining step components $\widehat{\Delta s}$ and $\widehat{\Delta x}$ are obtained by substituting into (2.15b) and (2.15c), respectively, where once again we assume that finite-precision arithmetic is used.

5.4. Checking assumptions and estimates for Theorem 4.2. We now prepare to apply Theorem 4.2 by checking that its various assumptions are satisfied for μ sufficiently small. We assume that ϵ is set to the following value:

$$\epsilon = 100\mathbf{u}. \tag{5.16}$$

This choice is motivated by a desire to keep ϵ as small as possible, while trying to ensure that the set $\mathcal{J}$ of skipped pivot indices is not greatly affected by the use of finite-precision arithmetic (see the discussion surrounding (4.8)). The assumption (3.10) that the largest diagonal in ρAD^2A^T is 1 is satisfied by construction. From (5.13) and (5.12), the assumptions (3.31) and (3.43) hold trivially. As noted in the discussion following (4.9), this assumption too can be expected to hold in nonpathological cases. It follows immediately from (4.9) that

$$\tau = \delta_1 \bar{\epsilon}^{-1/2}, \tag{5.17}$$

giving us a "worst-case" bound for τ. When **modchol** correctly identifies the numerical rank of AD^2A^T—that is, when $|\bar{\mathcal{J}}| = p = \operatorname{rank} A_{\cdot\mathcal{B}}$, as often happens in the examples we present in the next section—we usually have that all the diagonals of $\tilde{L}_{\bar{\mathcal{J}}\bar{\mathcal{J}}}$ are of size δ_1, and hence that $\tau = \delta_1$. Surprisingly, however, our favorable results about the quality of the computed step $(\widehat{\Delta x}, \widehat{\Delta \pi}, \widehat{\Delta s})$ hold even when the algorithm admits some small diagonal elements into $\tilde{L}_{\bar{\mathcal{J}}\bar{\mathcal{J}}}$, yielding a computed factor $\tilde{L}_{\bar{\mathcal{J}}\bar{\mathcal{J}}}$ for which $|\bar{\mathcal{J}}| > p$.

Having verified that we can reasonably expect Theorem 4.2 to hold for the system (5.15), we now estimate the quantities on the right-hand side of the bound (4.10). From (5.13a), we have $\sigma_p^{-1} = \Theta(1)$, while from (5.13b), we have $\sigma_{p+1} = \Theta(\mu)$.

We need to account, too, for the errors incurred in finite-precision evaluation of the right-hand side of (5.15), and to apportion these errors between the error vectors f and e in (4.3). For the purpose of this discussion, and in the remainder of the paper, we assume that

$$\mu \geq \mathbf{u}. \tag{5.18}$$

(As we see later, the algorithm is usually terminated—and for good reason—when μ is significantly larger than $\mathbf{u}$, so this assumption is not restrictive.) We examine the contributions of the terms r_{xs}, r_b, and r_c to the right-hand side of (5.15) in turn.

In most codes, the contribution of r_{xs} to (5.15) is calculated by forming the vector r_{xs}, multiplying by $D^2X^{-1} = S^{-1}$, and then multiplying by A. Floating-point error in formation of r_{xs} from (5.3) can be bounded by a term of size $\delta_{\mathbf{u}}\mu$. This error is magnified to $\delta_{\mathbf{u}}$ when we multiply by S^{-1}, and further roundoff errors introduced in this operation result in an additional error of size $\delta_{\mathbf{u}}$. Multiplication by A yields additional errors of size $\delta_{\mathbf{u}}$. Therefore, the total contribution of this term to the error in the right-hand side of (5.15), after scaling by ρ, has magnitude $\delta_{\mathbf{u}}\mu$. We denote this error by e_{xs}; below, we include it in the unstructured error vector e in (4.3).

The vectors r_b and r_c both have size μ (see (5.1)), but they are calculated by summing and differencing real-number quantities of size δ_1, and hence incur cancellation error of size $\delta_{\mathbf{u}}$. We denote the calculated versions by $\hat{r}_b$ and $\hat{r}_c$, respectively, so that

$$\hat{r}_b - r_b = \delta_{\mathbf{u}}, \qquad \hat{r}_c - r_c = \delta_{\mathbf{u}}. \tag{5.19}$$

The contribution of the error in $\hat{r}_b$ to the right-hand side of (5.15) is small. After scaling by ρ, it contributes an error of size $\mu\delta_{\mathbf{u}}$, which we denote by e_b and incorporate into e.

The term involving r_c requires more careful consideration. Note from (5.1) and (5.19) that $\hat{r}_c = O(\mu) + \delta_{\mathbf{u}}$. When we multiply $\hat{r}_c$ by D^2, some of whose diagonal elements have size $\Theta(\mu^{-1})$, we incur additional error of $\delta_{\mathbf{u}}\mu^{-1}(\mu+\delta_{\mathbf{u}})$, which is equivalent to $\delta_{\mathbf{u}}$ because of (5.18). Therefore, we have

$$\text{comp}(D^2\hat{r}_c) = D^2(r_c + \delta_{\mathbf{u}}) + \delta_{\mathbf{u}} = D^2 r_c + D^2(\hat{r}_c - r_c) + \delta_{\mathbf{u}},$$

which has size δ_1. Finally, on multiplying by A, we incur additional roundoff error of $\delta_{\mathbf{u}}$, so in summary we have

$$\text{comp}(AD^2\hat{r}_c) = AD^2 r_c + AD^2(\hat{r}_c - r_c) + \delta_{\mathbf{u}}. \tag{5.20}$$

From (5.14), we have that

$$AD = \rho^{-1/2} L Q^T \tag{5.21}$$

for some orthogonal matrix Q, so by defining

$$f = \rho^{1/2} Q D(\hat{r}_c - r_c) = O(\mu^{1/2}) O(\mu^{-1/2}) \delta_{\mathbf{u}} = \delta_{\mathbf{u}}, \tag{5.22}$$

we have that

$$\rho A D^2(\hat{r}_c - r_c) = \rho^{1/2} L Q^T D(\hat{r}_c - r_c) = L^T f.$$

Hence, from (5.20), we see that the computed version of the term $\rho AD^2 r_c$ on the right-hand side of (5.15) differs from the exact quantity by $Lf + e_c$, where f is defined as in (5.22) and $e_c = \mu\delta_{\mathbf{u}}$. By adding the unstructured error contributions from the three right-hand-side terms in (5.15), we find that

$$e = e_{xs} + e_b + e_c = \mu\delta_{\mathbf{u}}. \tag{5.23}$$

We have pointed out already (see (5.4)) that $\Delta\pi = O(\mu)$. The one remaining important quantity on the right-hand side of (4.10) is $\|\widehat{\Delta\pi}\|$. By making further assumptions on the relative sizes of τ, $\mathbf{u}$, and ϵ, we can bound this term strictly in terms of $\|\Delta\pi\|$, but the resulting estimate is observed to be too pessimistic. We found the following estimate to hold in all computational tests we performed:

$$\widehat{\Delta\pi} = O(\mu); \tag{5.24}$$

we use this estimate in the results below.

5.5. Errors in the computed step and their consequences. We now have all the estimates needed to apply Theorem 4.2 to (5.15). By substituting $z = \Delta\pi$ and $\hat{z} = \widehat{\Delta\pi}$, together with the estimates (5.13), (5.4), (5.24), (5.22), and (5.23), into (4.10), we obtain

$$\begin{aligned} \|L^T(\widehat{\Delta\pi} - \Delta\pi)\| &\le \delta_1 \left[(\tau\mathbf{u} + \bar{\epsilon}^{1/2}) + \mu\right]\mu + \delta_1(\bar{\epsilon}^{1/2} + \mu)\mu + (1 + \tau\mu)\delta_{\mathbf{u}} + \tau\mu\delta_{\mathbf{u}} \\ &= \delta_1\mu\left[\tau\mathbf{u} + \bar{\epsilon}^{1/2} + \mu + \mu^{-1}\mathbf{u}\right], \end{aligned} \tag{5.25}$$

and by substituting for τ from (5.17), we obtain

$$\|L^T(\widehat{\Delta\pi} - \Delta\pi)\| \le \delta_1\mu\left[\bar{\epsilon}^{-1/2}\mathbf{u} + \bar{\epsilon}^{1/2} + \mu + \mu^{-1}\mathbf{u}\right]. \tag{5.26}$$

From (5.21), and using orthogonality of Q, we can define

$$v = DA^T(\widehat{\Delta\pi} - \Delta\pi) \tag{5.27}$$

and note from (5.26) that

$$\|v\| = \rho^{-1/2}\|L^T(\widehat{\Delta\pi} - \Delta\pi)\| \leq \delta_1\mu^{1/2}\left[\bar{\epsilon}^{-1/2}\mathbf{u} + \bar{\epsilon}^{1/2} + \mu + \mu^{-1}\mathbf{u}\right]. \tag{5.28}$$

From (1.1) and (5.16), we see that the right-hand side of this expression is minimized, with a value of $\delta_1\mathbf{u}^{1/2}$, when $\mu \approx \bar{\epsilon}^{1/2} = \delta_1\mathbf{u}^{1/2}$. This observation suggests that a termination criterion of

$$\mu \leq \mathbf{u}^{1/2} \tag{5.29}$$

may be appropriate for the interior-point method. We justify this choice further below, after investigating the errors in the computed step and their effects on maximum steplength and on the updating of the residuals r_c and r_b.

Next, we examine the effect of the error in $\widehat{\Delta\pi}$ and the evaluation error in the right-hand side of (2.15b) on the calculated step $\widehat{\Delta s}$. From (5.4) and (5.24), we have that

$$\|\Delta\pi - \widehat{\Delta\pi}\| \leq \|\Delta\pi\| + \|\widehat{\Delta\pi}\| = O(\mu). \tag{5.30}$$

The evaluation error of size $\delta_{\mathbf{u}}$ in the r_c term of (2.15b) (see (5.19)) is significant; the additional roundoff errors of size $\mu\delta_{\mathbf{u}}$ incurred in forming the matrix–vector product and in performing the vector addition to evaluate the right-hand side of (2.15b) are negligible. We conclude from (5.19) and (5.30) that the computed step $\widehat{\Delta s}$ and exact step Δs differ as follows:

$$\Delta s - \widehat{\Delta s} = -r_c + \hat{r}_c - A^T(\Delta\pi - \widehat{\Delta\pi}) + \mu\delta_{\mathbf{u}} = \delta_1(\mu + \mathbf{u}). \tag{5.31}$$

This estimate is potentially troubling: Since the exact step Δs has size $O(\mu)$, it indicates that the computed step $\widehat{\Delta s}$ may not be correct to any digits at all! This inaccuracy is not so important for the "large" components of s—namely, components in the subvector $s_{\mathcal{N}}$—since eventually μ is small in comparison to these components and errors in $\Delta s_{\mathcal{N}}$ have little effect on the steplength α or on the updated value of x^Ts. However, errors of the size indicated in (5.31) in the $\mathcal{B}$ components of Δs could be disastrous. The consequences could include that the maximum steplength $\alpha_{\max}$ to the boundary could be much smaller than 1 (an argument similar to the one following (5.9) indicates only that $1 - \alpha_{\max} = \delta_1$) and in fact we cannot even be sure of decrease in x^Ts along this direction. Fortunately, a refined estimate of the error in $\widehat{\Delta s}_{\mathcal{B}}$ is possible. By using (5.19) in (5.31), we have that

$$\Delta s - \widehat{\Delta s} = -A^T(\Delta\pi - \widehat{\Delta\pi}) + \delta_{\mathbf{u}} = D^{-1}v + \delta_{\mathbf{u}}, \tag{5.32}$$

where v is defined as in (5.27). From (5.2), we have that $D_{ii}^{-1} = \Theta(\mu^{1/2})$ for $i \in \mathcal{B}$, and therefore, by using (5.28), we obtain

$$\Delta s_i - \widehat{\Delta s}_i = \delta_1\mu\left[\bar{\epsilon}^{-1/2}\mathbf{u} + \bar{\epsilon}^{1/2} + \mu + \mu^{-1}\mathbf{u}\right], \qquad i \in \mathcal{B}. \tag{5.33}$$

As in the discussion following (5.9), we find that $s_i + \alpha\widehat{\Delta s}_i = 0$ is possible only if

$$1 - \alpha = \delta_1\left[\bar{\epsilon}^{-1/2}\mathbf{u} + \bar{\epsilon}^{1/2} + \mu + \mu^{-1}\mathbf{u}\right]. \tag{5.34}$$

Finally, we estimate the errors in the computed step $\widehat{\Delta x}$ obtained from (2.15c) and estimate their effect on $\alpha_{\max}$ and on the updated value of r_b. Again, we consider the components $i \in \mathcal{B}$ and $i \in \mathcal{N}$ separately.

For $i \in \mathcal{B}$, the $\delta_{\mathbf{u}}\mu$ evaluation error in $(r_{xs})_i$ is magnified by the term $s_i^{-1} = \Theta(\mu^{-1})$. Floating-point error in forming the product $x_i\widehat{\Delta s}_i$ and in performing the addition yield additional errors of size at most $\delta_{\mathbf{u}}$, so we obtain

$$\Delta x_i - \widehat{\Delta x}_i = -s_i^{-1}x_i(\Delta s_i - \widehat{\Delta s}_i) + \delta_{\mathbf{u}}, \qquad i \in \mathcal{B}. \tag{5.35}$$

From (5.33) and (5.1), this formula implies that

$$\widehat{\Delta x}_i - \Delta x_i = \delta_1\left[\bar{\epsilon}^{-1/2}\mathbf{u} + \bar{\epsilon}^{1/2} + \mu + \mu^{-1}\mathbf{u}\right], \qquad i \in \mathcal{B}. \tag{5.36}$$

By the usual reasoning, we find that $x_i + \alpha\widehat{\Delta x}_i = 0$ is possible for $i \in \mathcal{B}$ only for α satisfying (5.34).

For $i \in \mathcal{N}$, the $\delta_{\mathbf{u}}\mu$ evaluation error in $(r_{xs})_i$ is not magnified appreciably by the term s_i^{-1} (which has size $\Theta(1)$), and we obtain

$$\Delta x_i - \widehat{\Delta x}_i = -s_i^{-1}x_i(\Delta s_i - \widehat{\Delta s}_i) + \mu\delta_{\mathbf{u}}, \qquad i \in \mathcal{N}. \tag{5.37}$$

By substituting from (5.31) and (5.1), we obtain

$$\widehat{\Delta x}_i - \Delta x_i = \delta_1[\mu^2 + \mu\mathbf{u}], \qquad i \in \mathcal{N}. \tag{5.38}$$

We deduce that $x_i + \alpha\widehat{\Delta x}_i = 0$ for $i \in \mathcal{N}$ only if

$$1 - \alpha = \delta_1[\mu + \mathbf{u}]. \tag{5.39}$$

From (5.34) and (5.39), we conclude that the value of $\alpha_{\max}$ defined by (2.13), with the calculated direction $(\widehat{\Delta x}, \widehat{\Delta \pi}, \widehat{\Delta s})$ replacing the exact search direction, satisfies the estimate

$$1 - \alpha_{\max} = \delta_1\left[\bar{\epsilon}^{-1/2}\mathbf{u} + \bar{\epsilon}^{1/2} + \mu + \mu^{-1}\mathbf{u}\right]. \tag{5.40}$$

Note from (5.30), (5.31), and (5.38) that, in an *absolute* sense, the errors in $\widehat{\Delta \pi}$, $\widehat{\Delta s}$, and $\widehat{\Delta x}_{\mathcal{N}}$ are small. By contrast, the $\mu^{-1}\mathbf{u}$ term in (5.36) implies that the errors in $\widehat{\Delta x}_{\mathcal{B}}$ increase as μ decreases below $\mathbf{u}^{1/2}$. These errors have consequences for the updated values of the residuals r_b and r_c at the new point

$$(x, \pi, s) + \alpha(\widehat{\Delta x}, \widehat{\Delta \pi}, \widehat{\Delta s}),$$

where $\alpha \in (0, \alpha_{\max})$ is the steplength chosen by the algorithm. From (2.7), we see that the computed value of r_c at this new point is given by

$$\mathrm{comp}(\hat{r}_c^+) = A^T(\pi + \alpha\widehat{\Delta \pi}) + (s + \alpha\widehat{\Delta s}) - c + \delta_{\mathbf{u}},$$

where the final term accounts for both cancellation and roundoff errors. From (5.32), we see that this quantity differs from the exact value of r_c^+ by

$$\alpha A^T(\widehat{\Delta \pi} - \Delta\pi) + \alpha(\widehat{\Delta s} - \Delta s) + \delta_{\mathbf{u}} = \delta_{\mathbf{u}},$$

so we conclude that the effect of the errors in $(\widehat{\Delta x}, \widehat{\Delta \pi}, \widehat{\Delta s})$ on the r_c term is minimal (that is, it is of the same order as the cancellation error that arises in any case when this term is evaluated).

The computed version of r_b at the new point is

$$\text{comp}(\hat{r}_b^+) = A(x + \alpha \widehat{\Delta x}) - b + \delta_{\mathbf{u}},$$

which differs from the exact version r_b^+ as follows:

$$\text{comp}(\hat{r}_b^+) - r_b^+ = \alpha A(\widehat{\Delta x} - \Delta x) + \delta_{\mathbf{u}}.$$

By substituting from (5.35) and (5.37) and using (2.14), we obtain

$$\text{comp}(\hat{r}_b^+) - r_b^+ = \alpha A D^2 (\Delta s - \widehat{\Delta s}) + \delta_{\mathbf{u}},$$

which, from (5.19) and (5.31) and the estimate $\|D^2\| = O(\mu^{-1})$, yields

$$\text{comp}(\hat{r}_b^+) - r_b^+ = \alpha A D^2 A^T (\widehat{\Delta \pi} - \Delta \pi) + \mu^{-1} \delta_{\mathbf{u}}. \tag{5.41}$$

From (5.21), (3.4), and (3.6), we have that

$$AD^2A^T = \rho^{-1} L L^T = \rho^{-1} (\tilde{L}\tilde{L}^T + E),$$

so by some elementary manipulation, we deduce that $\text{comp}(\hat{r}_b^+) - r_b^+$ equals the expression

$$\alpha\rho^{-1}\tilde{L}\tilde{L}^T(\tilde{\Delta \pi} - \Delta\pi) + \alpha\rho^{-1}E(\widehat{\Delta\pi} - \Delta\pi) + \alpha\rho^{-1}\tilde{L}\tilde{L}^T(\widehat{\Delta\pi} - \tilde{\Delta\pi}) + \mu^{-1}\delta_{\mathbf{u}}. \tag{5.42}$$

We bound this expression one term at a time, using results from earlier sections and identifying $\Delta\pi$ with z, $\widehat{\Delta\pi}$ with $\hat{z}$, and $\tilde{\Delta\pi}$ with $\tilde{z}$. For the first term, we have from (3.10) that $\|\tilde{L}\| \le \delta_1$, while from Theorem 3.4, (5.16), and (5.4), we have

$$\|\tilde{L}^T(\Delta\pi - \tilde{\Delta\pi})\| = \delta_{\mathbf{u}}^{1/2} \|\Delta\pi_{\mathcal{J}}\| = \mu\delta_{\mathbf{u}}^{1/2}. \tag{5.43}$$

For the second term in (5.42), we have from Lemma 3.2, (5.16), (5.30), and $\rho = \Theta(\mu)$ that

$$\rho^{-1}\|E(\widehat{\Delta\pi} - \Delta\pi)\| \le \delta_{\mathbf{u}}^{1/2}. \tag{5.44}$$

For the third term, recall that $\tilde{\Delta\pi}_{\mathcal{J}} = \widehat{\Delta\pi}_{\mathcal{J}} = 0$ and $\tilde{L}_{\cdot\mathcal{J}} = 0$, so that

$$\|\tilde{L}\tilde{L}^T(\tilde{\Delta\pi} - \widehat{\Delta\pi})\| \le \|\tilde{L}\| \|\tilde{L}^T_{\bar{\mathcal{J}}\bar{\mathcal{J}}}(\tilde{\Delta\pi}_{\bar{\mathcal{J}}} - \widehat{\Delta\pi}_{\bar{\mathcal{J}}})\| \le \delta_1 \|\tilde{L}^T_{\bar{\mathcal{J}}\bar{\mathcal{J}}}(\tilde{\Delta\pi}_{\bar{\mathcal{J}}} - \widehat{\Delta\pi}_{\bar{\mathcal{J}}})\|. \tag{5.45}$$

From (3.9), we have

$$\tilde{L}_{\bar{\mathcal{J}}\bar{\mathcal{J}}}\tilde{L}^T_{\bar{\mathcal{J}}\bar{\mathcal{J}}}\tilde{\Delta\pi}_{\bar{\mathcal{J}}} = r_{\bar{\mathcal{J}}},$$

and so from (3.5a), (4.3), and (4.4), we have

$$(\tilde{L}_{\bar{\mathcal{J}}\bar{\mathcal{J}}}\tilde{L}^T_{\bar{\mathcal{J}}\bar{\mathcal{J}}} + E^{\mathbf{u}}_{\bar{\mathcal{J}}\bar{\mathcal{J}}})\widehat{\Delta\pi}_{\bar{\mathcal{J}}} = \tilde{L}_{\bar{\mathcal{J}}\bar{\mathcal{J}}}\tilde{L}^T_{\bar{\mathcal{J}}\bar{\mathcal{J}}}\tilde{\Delta\pi}_{\bar{\mathcal{J}}} + (\tilde{L}\tilde{f} + \tilde{e})_{\bar{\mathcal{J}}}.$$

By rearranging, we obtain

$$\tilde{L}^T_{\bar{\mathcal{J}}\bar{\mathcal{J}}}(\widehat{\Delta\pi}_{\bar{\mathcal{J}}} - \tilde{\Delta\pi}_{\bar{\mathcal{J}}}) = -\tilde{L}^{-1}_{\bar{\mathcal{J}}\bar{\mathcal{J}}}\left[E^{\mathbf{u}}_{\bar{\mathcal{J}}\bar{\mathcal{J}}}\widehat{\Delta\pi}_{\bar{\mathcal{J}}} - \tilde{L}_{\bar{\mathcal{J}}\bar{\mathcal{J}}}\tilde{f}_{\bar{\mathcal{J}}} - \tilde{e}_{\bar{\mathcal{J}}}\right].$$

We now use the estimates

$$\begin{array}{rcll} \|\tilde{L}^{-1}_{\bar{\mathcal{J}}\bar{\mathcal{J}}}\| &=& \delta_{\mathbf{u}}^{-1/2} & \text{from (3.45), (5.16), and (5.17),} \\ \|E^{\mathbf{u}}_{\bar{\mathcal{J}}\bar{\mathcal{J}}}\| &=& \delta_{\mathbf{u}} & \text{from (4.2),} \\ \|f\| &=& \delta_{\mathbf{u}} & \text{from (4.5), (5.13a), and (5.22),} \\ \|\tilde{e}\| &=& \delta_{\mathbf{u}}^{3/2} + \mu\delta_{\mathbf{u}} & \text{from (4.5), (5.13), (5.22), and (5.23),} \\ \|\widehat{\Delta\pi}_{\bar{\mathcal{J}}}\| &=& O(\mu) & \text{from (5.24)} \end{array}$$

to yield the following bound:

$$\begin{aligned} \|\tilde{L}^T_{\bar{\mathcal{J}}\bar{\mathcal{J}}}(\widehat{\Delta\pi}_{\bar{\mathcal{J}}} - \Delta\pi_{\bar{\mathcal{J}}})\| &\le \|\tilde{L}^{-1}_{\bar{\mathcal{J}}\bar{\mathcal{J}}}\|\,\|E^{\mathbf{u}}_{\bar{\mathcal{J}}\bar{\mathcal{J}}}\|\,\|\widehat{\Delta\pi}_{\bar{\mathcal{J}}}\| + \|\tilde{f}_{\bar{\mathcal{J}}}\| + \|\tilde{L}^{-1}_{\bar{\mathcal{J}}\bar{\mathcal{J}}}\|\,\|\tilde{e}_{\bar{\mathcal{J}}}\| \\ &\le \delta_{\mathbf{u}}^{-1/2}\delta_{\mathbf{u}}\mu + \delta_{\mathbf{u}} + \delta_{\mathbf{u}}^{-1/2}[\delta_{\mathbf{u}}^{3/2} + \mu\delta_{\mathbf{u}}] \\ &\le \mu\delta_{\mathbf{u}}^{1/2} + \delta_{\mathbf{u}}. \end{aligned}$$

Therefore, for the third term in (5.42), we have from (5.45) that

$$\|\tilde{L}\tilde{L}^T(\tilde{\Delta\pi} - \widehat{\Delta\pi})\| \le \mu\delta_{\mathbf{u}}^{1/2} + \delta_{\mathbf{u}}. \tag{5.46}$$

By substituting (5.43), (5.44), (5.46), $\rho = \Theta(\mu)$, and $|\alpha| \le 1$ into (5.42), we have

$$\operatorname{comp}(\hat{r}_b^+) - r_b^+ = \delta_{\mathbf{u}}^{1/2} + \mu^{-1}\delta_{\mathbf{u}}. \tag{5.47}$$

This estimate suggests that the discrepancy between $\hat{r}_b^+$ and its approximation $\operatorname{comp}(\hat{r}_b^+)$ is no greater than $\delta_{\mathbf{u}}^{1/2}$ until μ falls below approximately $\mathbf{u}^{1/2}$. This observation, together with (5.40), suggests strongly that the termination condition (5.29) is the appropriate one. These observations too are illustrated in section 6.

The convergence tolerances used by most interior-point codes—arrived at by practical experience rather than theoretical or analytical considerations—are generally consistent with (5.29). For instance, the code PCx declares optimality if the following three conditions are satisfied:

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

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.

5.6. Comments and observations. We conclude this section with a few comments about the results above.

Note first that our conclusions can always be defeated by poor scaling of the problem. Poor scaling may show up as imbalance in the size of the components of $x_{\mathcal{B}}$ or $s_{\mathcal{N}}$ (some may be much smaller than others) or as imbalance in the sizes of the nonzero components of the problem data A, b, and c. Difficulties such as these may cause the many factors δ_1 that appear in the analysis to actually be much larger than 1, thereby limiting the regime of applicability of our results and affecting our conclusions about appropriate choices of $\bar{\epsilon}$ and the termination criterion. Most interior-point codes try to avoid these potential difficulties by prescaling the matrix A by some heuristic procedures, for example, the one proposed by Curtis and Reid [2].

A second point concerns the matrix $A_{\cdot\mathcal{B}}$, the basic part of the constraint matrix A. Our analysis is quite general in that it allows $A_{\cdot\mathcal{B}}$ to be rank deficient. However, when the nonzero singular values of this matrix are widely separated, the assumed separation (5.13) between the $p = \operatorname{rank} A_{\cdot\mathcal{B}}$ largest and $m - p$ smallest eigenvalues of AD^2A^T will not appear until μ is very small. This may again limit the regime of applicability of our analysis. Prescaling of the matrix A may help but, in some sense, ill conditioning of this type is intrinsic to the problem. As in many other areas of numerical linear algebra, it is not possible to design algorithms that produce accurate results in finite-precision arithmetic regardless of the conditioning of the problem.

Third, we note that our analysis made no assumption to ensure that **modchol** eventually determines the numerical rank of AD^2A^T. That is, none of our results require that $|\bar{\mathcal{J}}| = p$ for all μ sufficiently small. Although we observed that $|\bar{\mathcal{J}}| = p$ in many numerical tests, the assumptions needed to guarantee this equality are not satisfying in certain respects. (Such assumptions did appear in an earlier version of this paper, but they were discarded.) The advantage of $|\bar{\mathcal{J}}| = p$ in the analysis is that the matrix $\tilde{L}_{\bar{\mathcal{J}}\bar{\mathcal{J}}}$ will have all its diagonal elements of size $\Theta(1)$, allowing us to use the estimate $\tau = \delta_1$ in place of the weaker estimate (5.17). This estimate in turn allows us to bound the norm $\|\hat{z}\|$ in (4.10) in terms of $\|z\|$, leading to a more rigorous bound on $\|L^T(\hat{z} - z)\|$.

A fourth, related point concerns our estimate (5.17) of the size of τ, which is based on the assumption that the norm of $\tilde{L}^{-1}_{\bar{\mathcal{J}}\bar{\mathcal{J}}}$ can be estimated accurately by observing the sizes of its diagonal elements. While the resulting estimate appears to hold for the vast majority of practical problems of the type in question, there are cases in which it underestimates the value of $\|\tilde{L}^{-1}_{\bar{\mathcal{J}}\bar{\mathcal{J}}}\|$. See Lawson and Hanson [8, p. 31] for a classic example.

Finally, we note that when all the skipped pivots occur in the lower right corner of the matrix M (as happens on most of the smaller problems we tested), we can replace the bound $\|E\| \le \bar{\epsilon}^{1/2}$ by the tighter bound $\|E\| \le \bar{\epsilon}$. This tighter estimate allows some of our results to be strengthened, but since we observed some large linear programs in which the skipped pivots were not confined to the lower right corner, we omit a detailed analysis of this case.

6. Implementation and computational results. The **modchol** approach can be implemented by making minimal changes to a standard sparse Cholesky code. We need to 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. The pivot skipping itself can be performed explicitly (by inserting a column of zeros in the Cholesky factor and maintaining a record of the set $\mathcal{J}$), or it can be "simulated," as in LIPSOL [20] and PCx [3], by inserting a huge element in the pivot position prior to the computation of the column of the Cholesky factor and updating of the remainder of the matrix. In PCx [3], we needed to change fewer than 20 lines of the sparse Cholesky code of Ng and Peyton [10].

To test that the analysis of this paper was reflected in computations, we coded a simple primal-dual interior-point algorithm and applied it to test problems with controlled degeneracy properties. At each iterate, we monitored various quantities, compared them against the estimates of section 5, and confirmed that convergence to a tolerance of approximately $\mathbf{u}^{1/2}$ could be attained even for difficult problems.

Our test problems have the form (2.1), with $m = 6$ and $n = 12$. The matrix A is fully dense, with elements $(\xi_1 - .5)10^{6(\xi_2 - .5)}$, where ξ_1 and ξ_2 are random variables drawn from a uniform distribution on the interval $[0, 1]$. (Of course, the values of ξ_1 and ξ_2 are different for each element of the matrix.) After fixing the number of indices to appear in $\mathcal{B}$, we set

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

(Note that the problem is degenerate whenever $|\mathcal{B}| \neq 6$.) A primal solution x^* is constructed with

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

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

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

where ξ is random as above. Finally, we set $b = Ax^*$ and $c = A^T\pi^* + s^*$. Note that by our choice of $\mathcal{B}$, $A_{\cdot\mathcal{B}}$ consists of the last $|\mathcal{B}|$ columns of A. We modified A in some of the problems to introduce various types of rank deficiency.

The code was an implementation of the infeasible interior-point algorithm described by S. J. Wright [16]. The details of this algorithm are unimportant; we need note only that its iterates satisfy the estimates (5.1) in exact arithmetic and that the algorithm takes steps along the affine-scaling direction during its later iterations, provided that these steps make reasonable progress. At each iteration of the algorithm, we calculated the affine-scaling direction (whether or not it was actually used as a search direction) and kept a log of information about this step and about various other properties of the iterates and the **modchol** procedure. The parameter ϵ was set to 10^{-13}, which is about $500\mathbf{u}$ on the SPARCstation 5 that was used for the experiments. The results were not particularly sensitive to this parameter.

Results for various problems are shown in Tables 1, 2, 3, 4, and 5. For each iteration, we tabulate the norms $\|\widehat{\Delta x}^{\text{aff}}\|_\infty$, $\|\widehat{\Delta \pi}^{\text{aff}}\|_\infty$, and $\|\widehat{\Delta s}^{\text{aff}}\|_\infty$ of the affine-scaling step calculated at that iterate, together with the duality measure μ and residual norm $\|(r_b, r_c)\|_\infty$ for that iterate. We also tabulate the number of small pivots encountered during the factorization, that is, the number of elements in $\mathcal{J}$. The step-to-boundary $\alpha_{\max}$ along the calculated affine-scaling direction is also tabulated. (The algorithm actually uses the affine-scaling direction if this parameter exceeds 0.8; otherwise, it uses a direction with a centering component.) A horizontal line in each table indicates the iterate at which termination would occur if we use the termination criterion of section 5.5.

In Table 1 we chose $|\mathcal{B}| = m = 6$, making the linear program nondegenerate and the primal-dual solution unique. Note that the pivot-skipping mechanism in **modchol** is not activated for this problem, since the matrix AD^2A^T is approaching a well-conditioned limit. It is clear from the table that $\widehat{\Delta \pi}^{\text{aff}}$ and $\widehat{\Delta s}^{\text{aff}}$ satisfy the estimates (5.24) and (5.31), respectively, even when the algorithm continues to iterate 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 (5.36) when μ falls below $\mathbf{u}$. As discussed in section 5.5, this error is transmitted to the computed residual r_b, destroying the quality of subsequent iterates. A similar deterioration is noted in the steplength $\alpha_{\max}$. These observations show that it is important for the interior-point algorithm to save the best iterate obtained so far, so that it can report this value if it happens to push beyond the appropriate point of termination.

Table 2 shows results for the case of a problem in which $|\mathcal{B}| = 4$ with $A_{\cdot\mathcal{B}}$ full rank, which causes the coefficient matrix in (2.15a) to have four eigenvalues of size $\Theta(\mu^{-1})$ and the remaining two of size $\Theta(\mu)$. The second column shows that **modchol** detects small pivots when μ becomes sufficiently small, and confirms that the quality

TABLE 1
Affine-scaling step properties for a problem with $m = 6$, $n = 12$, $|\mathcal{B}| = 6$, $\operatorname{rank} A_{\cdot\mathcal{B}} = 6$. $\|\cdot\| = \|\cdot\|_\infty$, *and the horizontal line represents the normal point of termination.*

Iteration	Small pivots	$\log\mu$	log $\|(r_b, r_c)\|$	log $\|\widehat{\Delta x}^{\mathrm{aff}}\|$	log $\|\widehat{\Delta \pi}^{\mathrm{aff}}\|$	log $\|\widehat{\Delta s}^{\mathrm{aff}}\|$	$\alpha_{\max}$
$\vdots$							
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
$\vdots$							

TABLE 2
Affine-scaling step properties for a problem with $m = 6$, $n = 12$, $|\mathcal{B}| = 4$, $\operatorname{rank} A_{\cdot\mathcal{B}} = 4$. $\|\cdot\| = \|\cdot\|_\infty$, *and the horizontal line represents the normal point of termination.*

Iteration	Small pivots	$\log\mu$	log $\|(r_b, r_c)\|$	log $\|\widehat{\Delta x}^{\mathrm{aff}}\|$	log $\|\widehat{\Delta \pi}^{\mathrm{aff}}\|$	log $\|\widehat{\Delta s}^{\mathrm{aff}}\|$	$\alpha_{\max}$
$\vdots$							
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	−13.0	−12.5	−15.8	−14.2	>.99999
19	2	−25.1	−12.3	−1.5	−15.9	−13.7	>.99999
20	2	−29.7	1.2	6.7	−15.9	−13.3	.00016
$\vdots$							

of interior-point steps remains high after this point, at least until an accuracy of $\mathbf{u}^{1/2}$ is achieved. The behavior of the algorithm for very small values of μ—beyond the point of normal termination—is the same as that of Table 1.

The locations of the small pivots detected by **modchol** for the problem reported in Table 2 were at the bottom left of the matrix. We noted earlier that when this is the case, we have that the estimate $\|E\| \le \bar{\epsilon}^{1/2}$ of Lemma 3.2 can be replaced by the stronger estimate $\|E\| \le \bar{\epsilon}$. To show that the algorithm's performance does not depend critically on this smaller value of the error, we modified A to obtain a number of examples in which the small pivots appeared in locations other than the lower right of the matrix. In the problem report in Table 3, we modified the matrix A by replacing all elements in rows 1 and 2 with zeros, except for the element in the last column. We chose $|\mathcal{B}| = 6$, so that the matrix $A_{\cdot\mathcal{B}}$ formed by the last 6 columns of A has rank 5. Moreover, the fact that rows 1 and 2 of A are multiples of each other ensures that the $(2, 2)$ pivot will be flagged as a small pivot in **modchol**. It also implies that the

Table 3

Affine-scaling step characteristics for a problem with $m = 6$, $n = 12$, $|\mathcal{B}| = 6$, $\operatorname{rank} A_{\cdot\mathcal{B}} = 5$ *(rows* 1 *and* 2 *of* A *have a single nonzero each, in the same column location).* $\|\cdot\| = \|\cdot\|_\infty$*, and the horizontal line represents the normal point of termination.*

Iteration	Small pivots	$\log \mu$	log $\|(r_b, r_c)\|$	log $\|\widehat{\Delta x}^{\text{aff}}\|$	log $\|\widehat{\Delta \pi}^{\text{aff}}\|$	log $\|\widehat{\Delta s}^{\text{aff}}\|$	$\alpha_{\max}$
⋮							
11	1	−0.5	−12.6	0.3	1.6	0.8	.23771
12	1	−1.2	−10.3	0.6	1.0	0.2	.81949
13	1	−1.9	−10.3	0.9	0.1	−0.7	.67937
14	1	−2.4	−10.2	1.0	−0.9	−1.7	.50171
15	1	−3.4	−10.2	0.0	−2.3	−3.0	.95044
16	1	−4.7	−9.7	−1.0	−5.0	−5.0	.99199
17	1	−6.8	−11.3	−3.1	−7.1	−7.1	.99991
18	1	−10.9	−10.4	−0.3	−11.2	−11.1	.90487
19	1	−11.9	−10.3	0.3	−12.3	−12.2	.53423
⋮							

Table 4

Affine-scaling step characteristics for a problem with $m = 6$, $n = 12$, $|\mathcal{B}| = 4$, $\operatorname{rank} A_{\cdot\mathcal{B}} = 3$ *(*$A_{\cdot\mathcal{B}}$ *has two dependent columns).* $\|\cdot\| = \|\cdot\|_\infty$*, and the horizontal line represents the normal point of termination.*

Iteration	Small pivots	$\log \mu$	log $\|(r_b, r_c)\|$	log $\|\widehat{\Delta x}^{\text{aff}}\|$	log $\|\widehat{\Delta \pi}^{\text{aff}}\|$	log $\|\widehat{\Delta s}^{\text{aff}}\|$	$\alpha_{\max}$
⋮							
11	0	−0.4	−12.5	0.2	−0.4	1.1	.86945
12	0	−1.3	−11.2	−0.9	0.6	2.5	.19214
13	0	−1.8	−9.3	−0.9	−3.4	−1.5	>.99999
14	0	−3.8	−11.9	−3.2	−2.3	−0.4	.99848
15	3	−6.7	−9.5	−5.0	−8.0	−6.1	.99999
16	3	−11.8	−12.5	−0.2	−13.1	−11.1	.98866
17	3	−13.8	−12.6	1.9	−13.8	−11.9	.85592
18	3	−14.7	−13.5	−5.3	−13.2	−11.3	.92960
19	3	−15.8	−6.5	−6.5	−13.7	−11.7	>.99999
⋮							

assumption that A has full rank is violated. Table 3 confirms that the quality of the interior-point steps remains high. The algorithm's behavior is qualitatively the same as in the earlier examples.

The results in Table 3 illustrate that, as predicted by the analysis, the use of **modchol** does not cause the interior-point algorithm to break down even when $A_{\cdot\mathcal{B}}$ is rank deficient. We confirm this observation in Tables 4 and 5 with two other experiments involving rank-deficient matrices. Table 4 reports a problem identical to that of Table 2 except that in the matrix A, the third-last column was replaced by a multiple of the second-last column. The matrices A and $A_{\cdot\mathcal{B}}$ are thereby rank deficient. When μ becomes sufficiently small, **modchol** detects a numerical rank of 3 in the matrix of (2.15a), and the interior-point algorithm behaves similarly to that in the earlier tables. In Table 5, the modifications of A used in Tables 3 and 4 were both performed, giving a matrix $A_{\cdot\mathcal{B}}$ of rank 3 such that the pivots are not all confined to the lower right corner of the matrix in (2.15a). (The $(2, 2)$ pivot is always small.) The behavior

TABLE 5

Affine-scaling step characteristics for a problem with $m = 6$, $n = 12$, $|\mathcal{B}| = 4$, $\operatorname{rank} A_{\cdot\mathcal{B}} = 3$ ($A_{\cdot\mathcal{B}}$ has two dependent columns, and the first two rows of A contain a single nonzero each, in the same column location). $\|\cdot\| = \|\cdot\|_\infty$, and the horizontal line represents the normal point of termination.

Iteration	Small pivots	$\log\mu$	log $\|(r_b, r_c)\|$	log $\|\widehat{\Delta x}^{\text{aff}}\|$	log $\|\widehat{\Delta \pi}^{\text{aff}}\|$	log $\|\widehat{\Delta s}^{\text{aff}}\|$	$\alpha_{\max}$
$\vdots$							
11	1	−0.7	−10.0	0.3	2.9	2.4	.82144
12	1	−1.4	−9.3	−0.1	2.2	1.7	.85477
13	1	−2.2	−8.6	−0.5	−1.1	0.6	.50951
14	1	−2.5	−9.0	−0.8	−2.9	−1.3	.70461
15	2	−4.5	−10.5	−3.3	−2.0	−1.2	.99889
16	3	−7.5	−6.8	−5.4	−6.2	−4.2	>.99999
17	3	−12.9	−12.1	0.4	−11.9	−9.9	.95922
18	3	−14.3	−12.6	2.0	−13.3	−11.3	.20762
$\vdots$							

is once again similar to that of the earlier tables. We note especially iteration 15, at which two pivots are classified as "small" while a third pivot is slightly greater than the threshold, giving rise to a large spread in the nonzero diagonal elements of $\tilde{L}$. The resulting iterate contains some inaccuracy that manifests itself in a slight increase in the residual r_b, but this is quickly corrected at iteration 16, at which the large and small pivots become clearly separated.

Finally, we note that we tried degenerate test problems in which $|\mathcal{B}| > m$. These are less interesting because **modchol** detects no small pivots in factoring the matrix of (2.15a). Their behavior is once again similar to that of the other test problems, so we omit the details.

Acknowledgments. I am most grateful to the editor and referees. Their constructive comments on the first version of the paper led to considerable improvements, and their extremely close and patient reading of both versions led to the elimination of many infelicities and typos.

REFERENCES

[1] E. D. ANDERSEN AND K. D. ANDERSEN, *The MOSEK interior point optimizer for interior programming: An implementation of the homogeneous algorithm*, in High Performance Optimization Techniques, Kluwer Academic Publishers, Norwell, MA, to appear.

[2] A. R. CURTIS AND J. K. REID, *On the automatic scaling of matrices for Gaussian elimination*, J. Inst. Math. Appl., 10 (1972), pp. 118–124.

[3] 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, October 1996. Modified March 1997.

[4] I. I. DIKIN, *On the speed of an iterative process*, Upravlyaemye Sistemy, (1974).

[5] A. FORSGREN, P. GILL, AND J. R. SHINNERL, *Stability of symmetric ill-conditioned systems arising in interior methods for constrained optimization*, SIAM J. Matrix Anal. Appl., 17 (1996), pp. 187–211.

[6] J. GONDZIO, *HOPDM (version 2.12): A fast lp solver based on a primal-dual interior point method*, European J. Oper. Res., 85 (1995), pp. 221–225.

[7] N. J. HIGHAM, *Accuracy and Stability of Numerical Algorithms*, SIAM, Philadelphia, 1996.

[8] C. L. LAWSON AND R. J. HANSON, *Solving Least Squares Problems*, Prentice–Hall, Englewood Cliffs, NJ, 1974; reprinted as Classics in Appl. Math. 15, SIAM, Philadelphia, 1995.

[9] I. J. LUSTIG, R. E. MARSTEN, AND D. F. SHANNO, *Computational experience with a primal-*

dual interior point method for linear programming, Linear Algebra Appl., 152 (1991), pp. 191–222.

[10] E. Ng and B. W. Peyton, *Block sparse Cholesky algorithms on advanced uniprocessor computers*, SIAM J. Sci. Comput., 14 (1993), pp. 1034–1056.

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

[12] G. W. Stewart and J. Sun, *Matrix Perturbation Theory*, Comput. Sci. Sci. Comput., Academic Press, New York, 1990.

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

[14] R. J. Vanderbei and J. C. Lagarias, *Dikin's convergence result for the affine-scaling algorithm*, Contemp. Math., (1990).

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

[16] S. J. Wright, *A path-following interior-point algorithm for linear and quadratic optimization problems*, Ann. Oper. Res., 62 (1996), pp. 103–130.

[17] S. J. Wright, *Primal-Dual Interior-Point Methods*, SIAM, Philadelphia, 1997.

[18] S. J. Wright, *Stability of augmented system factorizations in interior-point methods*, SIAM J. Matrix Anal. Appl., 18 (1997), pp. 191–222.

[19] X. Xu, P. Hung, and Y. Ye, *A simplified homogeneous and self-dual linear programming algorithm and its implementation*, Ann. Oper. Res., 62 (1996), pp. 151–172.

[20] 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.