PREPRINT MCS-P446-0694, MATHEMATICS AND COMPUTER SCIENCE DIVISION, ARGONNE NATIONAL LABORATORY

Stability of Augmented System Factorizations in Interior-Point Methods

Stephen Wright*

July 7, 1995

Abstract

Some implementations of interior-point algorithms obtain their search directions by solving symmetric indefinite systems of linear equations. The conditioning of the coefficient matrices in these so-called augmented systems deteriorates on later iterations, as some of the diagonal elements grow without bound. Despite this apparent difficulty, the steps produced by standard factorization procedures are often accurate enough to allow the interior-point method to converge to high accuracy. When the underlying linear program is nondegenerate, we show that convergence to arbitrarily high accuracy occurs, at a rate that closely approximates the theory. We also explain and demonstrate what happens when the linear program is degenerate, where convergence to acceptable accuracy (but not arbitrarily high accuracy) is usually obtained.

1 Introduction

We focus on the core linear algebra operation in primal-dual interior-point methods for linear programming: solution of a system of linear equations whose coefficient matrix is large, sparse, and symmetric. In existing codes, the linear system is formulated in two different ways. One formulation, usually called the *augmented system formulation*, has a symmetric indefinite coefficient matrix. The other involves a more compact (but generally denser) symmetric positive-definite matrix. A diagonal matrix D is involved in both formulations, where D has the disconcerting property that some of its elements grow to ∞ as the iterates approach the solution set. This blowup in D can produce ill conditioning in the coefficient matrix of the linear system. In this paper, we examine the augmented system and look at how various factorization algorithms for this system behave as this ill conditioning develops.

We restrict our study to three standard factorization algorithms — the Bunch-Parlett, Bunch-Kaufman, and sparse Bunch-Parlett algorithms. The last of these has been used in at

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

least one practical interior-point code for linear programming (see Fourer and Mehrotra [4]). We assume that no attempt is made to improve the conditioning of the underlying linear systems by guessing whether each component of the solution is at a bound. Preprocessing of this kind detracts from the intuitive appeal of interior-point algorithms, namely, that they avoid explicit guessing about the contents of the basis.

In numerical experiments with feasible linear programs, we find that two distinct scenarios arise.

1. Even when the iterates are very close to the solution set, the computed search directions are good enough to produce rapid convergence of the algorithm at nearly the rates predicted by the theory. This performance is a little surprising. Since the matrix is poorly conditioned, we might have expected the computed directions to be too inaccurate to allow the algorithm to make much progress. This scenario usually occurs when the underlying linear program has a unique primal-dual solution.

2. Near the solution, calculation of the search direction fails because of breakdown of the matrix factorization, or else the computed search direction is so inaccurate that the interior-point method can move only a tiny distance along it before violating a bound. This scenario usually occurs when the underlying linear program is degenerate.

Our analysis in this paper explains these observations through a close examination of the behavior of factorization algorithms on the highly structured matrices that arise in our application. The effects of roundoff error are tracked by using fairly standard techniques from backward error analysis.

The most successful interior-point methods for practical linear programming problems are primal-dual methods. The best-known potential-reduction algorithm in this class was devised by Kojima, Mizuno, and Yoshise [8]; the review paper of Todd [16] contains a wealth of historical information on potential-reduction methods. Early developments in path-following methods are surveyed by Gonzaga [6], while Mizuno, Todd, and Ye [14] describe an important variant of these methods that does not require the iterates to stay within a cramped neighborhood of the central path. Zhang [24] extended the path-following approach further, allowing the iterates to be infeasible while retaining global convergence and polynomial complexity; see also Wright [20]. Some of these developments took place in the context of linear complementarity, a class of problems that includes linear programming as a special case.

On the computational side, the OB1 code described by Lustig, Marsten, and Shanno [9] generated search directions of the type described in this paper. They compute the maximum step α^* that could be taken along this direction without violating the positivity bounds, then set the actual step length to $.995\ \alpha^*$. Mehrotra's [13] predictor-corrector search direction differs from the one analyzed in this paper, but under our assumptions below, the difference vanishes as the solution is approached. Newer codes, such as those described by Mehrotra [13], Fourer and Mehrotra [4], Lustig, Marsten, and Shanno [11], Vanderbei [17], and Xu, Hung, and Ye [22] all implement Mehrotra's predictor-corrector strategy. These newer codes

continue to use step lengths based on α^*; hence, we pay particular attention to the effect of roundoff error on this quantity.

Previous analysis of the ill-conditioned linear systems that arise in interior-point and barrier methods has been carried out by Ponceleón [15] and Wright [21]. Ponceleón [15] showed that these systems are not too sensitive to structured perturbations from a certain class provided that the underlying optimization problem is well conditioned. Wright [21] analyzed Gaussian elimination in the context of interior-point algorithms for linear complementarity problems.

Simultaneously with the original version of this paper, and independently, Forsgren, Gill, and Shinnerl [3] performed an analysis of the augmented system in barrier algorithms. Their analysis tends to be more detailed than ours, and a few of the results overlap. However, they assume that the factorization algorithms select the large diagonal elements as pivots before any others, a pattern that does not generally occur in practice.

Vavasis [18] gives an illuminating discussion of the augmented system in other contexts besides optimization. He presents a solution method that is provably stable in a certain sense, but which is not guaranteed to produce "useful" steps in the sense of this paper. Duff [2] also discusses augmented systems in a general context and describes a sparse factorization procedure.

2 Interior-Point Methods

We consider the linear program in standard form:

$$\min c^T x, \qquad Ax = b, \qquad x \geq 0, \tag{1}$$

where $x \in \mathbb{R}^n$ and $b \in \mathbb{R}^m$. The dual of (1) is

$$\max b^T \lambda, \qquad A^T \lambda + s = c, \qquad s \geq 0, \tag{2}$$

where $s \in \mathbb{R}^n$ and $\lambda \in \mathbb{R}^m$. A vector triple (λ^*, x^*, s^*) is a primal-dual solution if x^* is feasible for (1), (λ^*, s^*) is feasible for (2), and s^* and x^* are complementary; that is,

$$x^{*T} s^* = c^T x^* - b^T \lambda^* = 0. \tag{3}$$

We denote the set of primal-dual solutions by $\mathcal{S}$.

Each iterate (λ, x, s) of a primal-dual interior-point method satisfies the strict inequality $(x, s) > 0$. Search directions are found by applying a modification of Newton's method to the following system of nonlinear equations:

$$Ax - b = 0, \qquad A^T \lambda + s - c = 0, \qquad XSe = 0, \tag{4}$$

where $X = \mathrm{diag}(x_1, x_2, \ldots, x_n)$ and $S = \mathrm{diag}(s_1, s_2, \ldots, s_n)$. Specifically, the search direction $(\Delta\lambda, \Delta x, \Delta s)$ satisfies the linear equations

$$\begin{bmatrix} 0 & A^T & I \\ A & 0 & 0 \\ S & 0 & X \end{bmatrix} \begin{bmatrix} \Delta x \\ \Delta \lambda \\ \Delta s \end{bmatrix} = \begin{bmatrix} -A^T\lambda - s + c \\ b - Ax \\ -XSe + \sigma\mu e \end{bmatrix}, \tag{5}$$

where $\sigma \in [0,1]$ is known as the centering parameter and the important quantity μ is defined by

$$\mu = x^T s/n.$$

The step length α along the search direction is determined by various factors; minimally, the updated x and s components are required to stay strictly positive:

$$(x,s) + \alpha(\Delta x, \Delta s) > 0. \tag{6}$$

At least half the components of (x,s) — the *critical* components — become very close to their lower bound of zero during the later stages of the algorithm. Despite this property, the step length α can be quite close to one without violating the property (6), when the search direction $(\Delta\lambda, \Delta x, \Delta s)$ is an *exact* solution of (5). If perturbations caused by roundoff are present in the critical components of $(\Delta\lambda, \Delta x, \Delta s)$, the requirement (6) can severely curtail the allowable step length and slow the convergence. Hence, it is important that the critical components of $(\Delta\lambda, \Delta x, \Delta s)$ be computed to high relative accuracy. This point provides the focus for much of our error analysis.

Throughout the paper we use $\mathbf{u}$ to denote unit roundoff, which we define implicitly by the statement that when x and y are any two floating-point numbers, op denotes $+, -, \times, /$, and $fl(z)$ denotes the floating-point approximation of any real number z, we have

$$fl(x \operatorname{op} y) = (x \operatorname{op} y)(1+\delta), \qquad |\delta| \le \mathbf{u}. \tag{7}$$

Since our concern is with the internal workings of a single interior-point iterate, we omit iteration counters from all quantities. For this reason, we use the order notation $O(\cdot)$ in a slightly unconventional way. When ξ and η are two nonnegative numbers, we write $\xi = O(\eta)$ if there is a positive constant C (not too large) such that $\xi \le C\eta$. We say that a matrix or vector is $O(\eta)$ if its norm is $O(\eta)$. We say that $\xi = \Omega(\eta)$ if $\xi = O(\eta)$ and $\eta = O(\xi)$.

For the purposes of this paper, we are mainly interested in how the factorizations behave relative to μ and $\mathbf{u}$. The dimensions m and n are ignored in our use of the notation $O(\cdot)$.

If G is a matrix, $G_{\cdot j}$ denotes its j-th column, while $G_{i\cdot}$ denotes the i-th row. The matrix whose elements are $|G_{ij}|$ is denoted by $|G|$.

We use $\|\cdot\|$ to denote any one of the equivalent matrix norms $\|\cdot\|_1$, $\|\cdot\|_2$, or $\|\cdot\|_\infty$. When G is rectangular, the 2-norm condition number is defined as follows.

Definition 1 *Let G be a rectangular matrix with full rank, and suppose that* $\mathbf{svmax}\ (G)$ *and* $\mathbf{svmin}\ (G)$ *denote the largest and smallest singular values of G, respectively. The 2-norm condition number of G is*

$$\kappa(G) = \frac{\mathbf{svmax}\ (G)}{\mathbf{svmin}\ (G)}.$$

If G is square and nonsingular, this definition coincides with the usual definition

$$\kappa(G) = \|G\|_2 \|G^{-1}\|_2.$$

3 Definitions and Assumptions

We assume throughout that the problems (1), (2) are feasible; that is, there exists at least one triple (λ, x, s) satisfying the constraints $Ax = b$, $A^T\lambda + s = c$, $(x, s) \geq 0$. Feasibility implies existence of solutions to (1), (2). The following theorem gives another consequence of feasibility.

Theorem 3.1 *Suppose that (1) and (2) are feasible and that (λ, x, s) is any point with $(x, s) > 0$. Then there exists a solution $(\Delta\lambda, \Delta x, \Delta s)$ to (5).*

Proof. The proof follows from Section 6 of Wright [20]. See, in particular, Lemma 6.2, Theorem 6.3, and the remarks in the last two paragraphs of [20]. ■

Note that A need not have full rank for Theorem 3.1 to hold.

The set of *basic* indices $\mathcal{B} \subset \{1, 2, \ldots, n\}$ can be defined as

$$\mathcal{B} = \{i \mid s_i^* = 0 \text{ for all } (\lambda^*, x^*, s^*) \in \mathcal{S}\}, \tag{8}$$

while the nonbasic set $\mathcal{N}$ is

$$\mathcal{N} = \{i \mid x_i^* = 0 \text{ for all } (\lambda^*, x^*, s^*) \in \mathcal{S}\}. \tag{9}$$

It is well known that $\mathcal{B}$ and $\mathcal{N}$ form a partition of $\{1, 2, \ldots, n\}$ and that there is at least one solution (λ^*, x^*, s^*) that is strictly complementary, that is, $x^* + s^* > 0$ (Goldman and Tucker [5]). The cardinality of $\mathcal{B}$ is denoted by $|\mathcal{B}|$. By partitioning the columns of A according to $\mathcal{B}$ and $\mathcal{N}$, we define

$$B = [A_{\cdot j}]_{j \in \mathcal{B}}, \qquad N = [A_{\cdot j}]_{j \in \mathcal{N}}, \tag{10}$$

so that B is $m \times |\mathcal{B}|$ and N is $m \times |\mathcal{N}|$. We say that the linear program is *nondegenerate* if $|\mathcal{B}| = m$ and the primal-dual solution is unique. We assume also that B is reasonably well conditioned in nondegenerate problems.

We do not confine our analysis to one specific primal-dual algorithm. Rather, we rely on a set of assumptions that is satisfied by a variety of algorithms. The first of these assumptions concerns the iterates, the search directions, and the relationship between μ and the current infeasibility.

Assumption 1 *The sequence of iterates (λ, x, s) generated by the interior-point algorithm satisfies the following properties when μ becomes sufficiently small:*

$$x_i = \Omega(1) \quad (i \in \mathcal{B}), \qquad s_i = \Omega(1) \quad (i \in \mathcal{N}), \tag{11a}$$

$$x_i = \Omega(\mu) \quad (i \in \mathcal{N}), \qquad s_i = \Omega(\mu) \quad (i \in \mathcal{B}). \tag{11b}$$

In addition, the infeasibility is $O(\mu)$; that is,

$$b - Ax = O(\mu), \qquad c - A^T\lambda - s = O(\mu). \tag{12}$$

Assumption 1 is not very strong. Güler and Ye [7] study algorithms in which all iterates are strictly feasible; that is

$$Ax = b, \qquad A^T\lambda + s = c, \qquad (x,s) > 0. \tag{13}$$

In fact they require that x and s be slightly separated from the boundary of the positive orthant, in the sense that

$$x_i s_i \geq \gamma\mu, \qquad i = 1, 2, \ldots, n, \tag{14}$$

for some constant $\gamma \in (0,1)$. They show that all limit points of such algorithms are strictly complementary solutions of (1), (2) and that most path-following and potential-reduction algorithms do in fact satisfy (14). It is easy to infer from their results that (11) holds for all subsequences that approach these limit points. Moreover, (12) is trivially satisfied for all feasible algorithms.

The infeasible-interior-point algorithm described by Wright [19] satisfies Assumption 1. So does the algorithm in [20], provided that the sequence or iterates (x,s) is bounded. Implemented algorithms such as those of Vanderbei [17], Lustig, Marsten, and Shanno [9, 10], and Xu, Hung, and Ye [22] usually step a fixed multiple of the distance to the boundary rather than enforce a potential reduction condition or a condition like (14). Nevertheless, the iteration sequence usually satisfies the properties of Assumption 1 for most practical problems.

Finally, we state without proof a technical lemma for use in later sections.

Lemma 3.2 *Let H be a square matrix partitioned as*

$$H = \begin{bmatrix} H_{11} & H_{12} \\ H_{21} & H_{22} \end{bmatrix},$$

where H_{11} and H_{22} are also square. Suppose that H_{11} and $H_{22} - H_{21}H_{11}^{-1}H_{12}$ are nonsingular. Then H is nonsingular, and

$$H^{-1} = \begin{bmatrix} H_{11}^{-1} + H_{11}^{-1}H_{12}(H_{22} - H_{21}H_{11}^{-1}H_{12})^{-1}H_{21}H_{11}^{-1} & -H_{11}^{-1}H_{12}(H_{22} - H_{21}H_{11}^{-1}H_{12})^{-1} \\ -(H_{22} - H_{21}H_{11}^{-1}H_{12})^{-1}H_{21}H_{11}^{-1} & (H_{22} - H_{21}H_{11}^{-1}H_{12})^{-1} \end{bmatrix}.$$

4 Exact and Approximate Search Directions

By defining $r_b = Ax - b$ and $r_c = A^T\lambda + s - c$ in (5), we obtain

$$\begin{bmatrix} 0 & A & 0 \\ A^T & 0 & I \\ 0 & S & X \end{bmatrix} \begin{bmatrix} \Delta\lambda \\ \Delta x \\ \Delta s \end{bmatrix} = \begin{bmatrix} -r_b \\ -r_c \\ -XSe + \sigma\mu e \end{bmatrix}. \tag{15}$$

By eliminating Δs from this system, we obtain the augmented system formulation:

$$\begin{bmatrix} 0 & A \\ A^T & -X^{-1}S \end{bmatrix} \begin{bmatrix} \Delta\lambda \\ \Delta x \end{bmatrix} = \begin{bmatrix} -r_b \\ -r_c + s - \sigma\mu X^{-1}e \end{bmatrix} \tag{16a}$$

$$\Delta s = -s + \sigma\mu X^{-1}e - X^{-1}S\Delta x. \tag{16b}$$

In Wright [21], we performed an error analysis on a system like (16a), but in the context of a specific path-following algorithm for the monotone linear complementarity problem. Some of our results from [21] are relevant to the present case of (16), as we discuss later.

Potential difficulties with the formulation (16) arise from two sources — possible rank-deficiency in certain submatrices of A, and the fact that some diagonal elements of $X^{-1}S$ and $S^{-1}X$ approach zero while others approach $+\infty$. Despite the effects of ill conditioning and finite precision, we find that the approximate search directions obtained from (16) by using standard factorization procedures are often remarkably good. They allow the interior-point algorithm to take near-unit steps and to make substantial improvements in the duality measure μ. In the following theorem, we specify a set of conditions for which this happy situation holds. In later sections, we identify situations under which these conditions hold.

In the remainder of the paper, we use α^* to denote the largest number in $[0,1]$ such that

$$(x+\alpha\Delta x, s+\alpha\Delta s) \geq 0 \qquad \text{for all } \alpha \in [0,\alpha^*]; \tag{17a}$$

$$(x+\alpha\Delta x)^T(s+\alpha\Delta s) \qquad \text{is decreasing for } \alpha \in [0,\alpha^*]. \tag{17b}$$

Theorem 4.1 *Suppose that Assumption 1 holds. Let $(\Delta\lambda, \Delta x, \Delta s)$ be the exact solution of (5) (equivalently, (16)), and let $(\widehat{\Delta\lambda}, \widehat{\Delta x}, \widehat{\Delta s})$ be an approximation to this step. Suppose that the centering parameter σ in (5) lies in the range $[0,1/2]$ and that the following conditions hold:*

$$(\Delta x, \Delta s) = O(\mu), \tag{18a}$$

$$(\Delta s_N, \Delta x_B) - (\widehat{\Delta s}_N, \widehat{\Delta x}_B) = O(\mathbf{u}), \tag{18b}$$

$$(\Delta s_B, \Delta x_N) - (\widehat{\Delta s}_B, \widehat{\Delta x}_N) = O(\mu\mathbf{u}). \tag{18c}$$

Define α^* as in (17), and suppose $\hat{\alpha}^*$ is obtained by replacing $(\Delta x, \Delta s)$ with $(\widehat{\Delta x}, \widehat{\Delta s})$ in (17). Then for all μ sufficiently small, we have

$$1-\alpha^* = O(\mu), \tag{19}$$

$$\hat{\alpha}^* = \alpha^* + O(\mathbf{u}) = 1 + O(\mu) + O(\mathbf{u}), \tag{20}$$

and

$$(x+\hat{\alpha}^*\widehat{\Delta x})^T(s+\hat{\alpha}^*\widehat{\Delta s})/n = \sigma O(\mu) + O(\mu(\mu+\mathbf{u})). \tag{21}$$

Proof. From (11a) and (18a), we have

$$s_N + \alpha\Delta s_N > 0, \qquad x_B + \alpha\Delta x_B > 0, \qquad \text{for all } \alpha \in [0,1],$$

so these components do not restrict the value of α^*. Since $\mathbf{u}$ is much smaller than 1, we use (18b) as well to deduce that

$$s_N + \alpha\widehat{\Delta s}_N \geq s_N + \alpha\Delta s_N + \alpha(\widehat{\Delta s}_N - \Delta s_N) > 0, \qquad \text{for all } \alpha \in [0,1].$$

Similarly, we can show that $x_B + \alpha\widehat{\Delta x}_B > 0$ for all $\alpha \in [0,1]$.

For the decrease condition (17b) we show that the duality gap actually decreases over the entire interval $[0,1]$ for both exact and approximate search directions, so that this condition does not play a role in determining α^* or $\hat{\alpha}^*$. For the exact direction, we have from (5), (18a), and $\sigma \in [0, 1/2]$ that

$$\begin{aligned} \frac{d}{d\alpha}(x+\alpha\Delta x)^T(s+\alpha\Delta s) &= x^T\Delta s + s^T\Delta x + 2\alpha\Delta x^T\Delta s \\ &\leq -(1-\sigma)n\mu + 2\|\Delta x\|\|\Delta s\| \\ &\leq -n\mu/2 + O(\mu^2), \end{aligned}$$

for all $\alpha \in [0,1]$. Hence, for μ sufficiently small, the duality gap is decreasing over $[0,1]$. For the approximate direction $(\widehat{\Delta x}, \widehat{\Delta s})$, this bound can be modified slightly to account for the inexactness. We omit the details, which are straightforward but messy, and state the conclusion as

$$\frac{d}{d\alpha}(x+\alpha\widehat{\Delta x})^T(s+\alpha\widehat{\Delta s}) \leq -n\mu/2 + O(\mu\mathbf{u} + \mu^2).$$

Again, we find that the duality gap is decreasing over the whole interval $\alpha \in [0,1]$.

Hence, the only condition that can bound α^* and $\hat{\alpha}^*$ away from 1 is (17a), and then only for the $\mathcal{N}$-components of x and the $\mathcal{B}$-components of s. In fact, α^* satisfies

$$\frac{1}{\alpha^*} = \max\left(1, \max_{i\in\mathcal{B}} -\frac{\Delta s_i}{s_i}, \max_{i\in\mathcal{N}} -\frac{\Delta x_i}{x_i}\right). \tag{22}$$

From (5), we have $x_i\Delta s_i + s_i\Delta x_i = -x_i s_i + \sigma\mu$. Hence, since $x_i s_i = \Omega(\mu)$ from (11), we have

$$-\frac{\Delta s_i}{s_i} = 1 + \frac{\Delta x_i}{x_i} - \sigma\frac{\mu}{x_i s_i} < 1 + \frac{\Delta x_i}{x_i}.$$

For $i \in \mathcal{B}$ we have from (11a) and (18a) that $|\Delta x_i/x_i| = O(\mu)$ and therefore

$$\max_{i\in\mathcal{B}} -\frac{\Delta s_i}{s_i} \leq 1 + O(\mu).$$

An identical argument can be used for the other term in (22), so we have

$$\frac{1}{\alpha^*} \leq \max(1, 1+O(\mu)) \quad \Rightarrow \quad 1-\alpha^* = O(\mu),$$

proving (19).

For the maximum step length $\hat{\alpha}^*$ along the approximate direction $(\widehat{\Delta x}, \widehat{\Delta s})$, we have from (18c) and (11b) that

$$\frac{\widehat{\Delta s}_i}{s_i} - \frac{\Delta s_i}{s_i} = \frac{O(\mu\mathbf{u})}{\Omega(\mu)} = O(\mathbf{u}), \quad (i \in \mathcal{B}), \qquad \frac{\widehat{\Delta x}_i}{x_i} - \frac{\Delta x_i}{x_i} = O(\mathbf{u}), \quad (i \in \mathcal{N}).$$

Hence, from (22), we have

$$\frac{1}{\hat{\alpha}^*} = \max\left(1, \max_{i\in\mathcal{B}} -\frac{\widehat{\Delta s}_i}{s_i}, \max_{i\in\mathcal{N}} -\frac{\widehat{\Delta x}_i}{x_i}\right) = \frac{1}{\alpha^*} + O(\mathbf{u}). \tag{23}$$

For all sufficiently small μ, the estimates (20) follow immediately from this last expression.

Finally, for the estimate of potential decrease (21), we have from (5) that

$$\begin{aligned}
&(x + \alpha\widehat{\Delta x})^T(s + \alpha\widehat{\Delta s}) \\
&= \left[x + \alpha\Delta x + \alpha(\widehat{\Delta x} - \Delta x)\right]^T \left[s + \alpha\Delta s + \alpha(\widehat{\Delta s} - \Delta s)\right] \\
&\leq n\mu(1 - \alpha(1-\sigma)) + O(\mu\mathbf{u}) + O(\mu\mathbf{u}^2),
\end{aligned} \tag{24}$$

where we have used Assumption 1 and (18) to estimate the remainder terms. Finally, we obtain (21) by substituting $\alpha = \hat{\alpha}^* = 1 + O(\mu + \mathbf{u})$ into (24). ■

5 The Augmented System

In the remainder of the paper, we focus on the procedure based on (16) for finding the search directions. In this section, we define a generalized form of the matrix in (16a) which we call a *canonical matrix*. We show that if the backward error analysis of the solution procedure satisfies a certain condition — Condition 1 below — then the approximate step $(\widehat{\Delta\lambda}, \widehat{\Delta x}, \widehat{\Delta s})$ obtained from (16) in a finite-precision environment is "useful" in the sense of Theorem 4.1.

In later sections, we define conditions under which these standard algorithms for solving symmetric indefinite systems satisfy Condition 1 and hence yield useful search directions. Our sharpened, specialized error analysis yields much stronger results than naive application of the standard results. We also gain insight into how the algorithms work even when the nondegeneracy assumptions of Sections 6, 7, and 8 fail to hold, and why they continue to generate useful search directions even for degenerate problems until μ is quite small.

Given a symmetric matrix T of order $\bar{n}$, the factorization procedures yield

$$LDL^T = PTP^T, \tag{25}$$

where P is a permutation matrix, L is unit lower triangular, and D is a block-diagonal matrix with 1×1 and 2×2 diagonal blocks. We denote the counterparts of these matrices that are actually computed in the finite-precision environment by $\hat{L}$ and $\hat{D}$, respectively.

Given the system $Tz = d$ and the data P, $\hat{L}$, and $\hat{D}$ from the factorization, we find the computed solution $\hat{z}$ by performing two vector permutations with P, triangular substitutions with $\hat{L}$ and $\hat{L}^T$, and a blockwise inversion of $\hat{D}$. Each of these operations (except the permutations) may introduce additional roundoff error, which must be accounted for in the error analysis.

For each of the methods, we focus on a single step of the factorization procedure applied to a matrix T with properties like those of our given system (16a), which we now define.

Definition 2 *A matrix T is a* canonical matrix *if it is a symmetric permutation of*

$$\begin{bmatrix} 0 & B & N \\ B^T & 0 & 0 \\ N^T & 0 & \Lambda \end{bmatrix} + O(\mu + \mathbf{u}), \tag{26}$$

where

- $\mu > 0$ *and* $\mathbf{u} \geq 0$ *are small;*
- Λ *is diagonal with all diagonal elements of magnitude* $\Omega(\mu^{-1})$*;*
- $B = \Omega(1)$ *and* $\kappa(B) = O(1)$*; and*
- $N = O(1)$.

We call T a degenerate canonical matrix *if it has the form*

$$\begin{bmatrix} 0 & 0 \\ 0 & \Lambda \end{bmatrix} + O(\mu + \mathbf{u}), \tag{27}$$

where the zero blocks are nonvacuous.

In keeping with our particular application (16a), we use m and n to denote the number of rows and columns in the composite matrix $[B \,|\, N]$, respectively, and $\bar{n} = m + n$ to denote the total dimension of T.

Corresponding to our canonical matrix, we define a canonical *error* matrix. We prove that for each of the factorizations, the error matrix has this form.

Definition 3 *Let T be a canonical matrix. The corresponding* canonical error matrix Δ *is a matrix of the same dimension as T such that*

$$|\Delta| \leq \Delta_{\mathbf{u}} + |T|\delta_{\mathbf{u}}, \tag{28}$$

where $\delta_{\mathbf{u}}$ and the elements of $\Delta_{\mathbf{u}}$ are $O(\mathbf{u})$.

An important role in the pivot selection process is played by the quantities χ_i, which denote the magnitude of the largest off-diagonal element in column i, that is,

$$\chi_i = \max\{|T_{ij}| \,|\, j = 1, 2, \ldots, \bar{n},\, j \neq i\}. \tag{29}$$

A sufficient condition for useful steps

The following condition states the common goal of our backward error analysis of the three factorization procedures. When this condition is satisfied along with nondegeneracy of the linear program, the result of Theorem 4.1 holds.

Condition 1 *Given the system $Tz = d$, where T is a canonical matrix, the symmetric factorization and solution process yields a computed solution $\hat{z}$ that satisfies*

$$(T + \Delta)\hat{z} = \hat{d}, \tag{30}$$

where Δ is a canonical error matrix associated with T and $\hat{d} - d = O(\mathbf{u})$.

We allow for a perturbed right-hand side $\hat{d}$ because of the nature of our particular system (16a). The residuals r_b and r_c are computed as the difference of $O(1)$ quantities, so $O(\mathbf{u})$ perturbations will appear when they are evaluated in the obvious way. Addition of the terms s_N and $\mu X_N^{-1} e$ may give rise to errors of similar magnitude.

Theorem 5.1 *Suppose that Assumption 1 holds and that the problem is nondegenerate, that is, $|\mathcal{B}| = m$, with $\kappa(B)$ moderate. Suppose that the procedure for solving (16) satisfies Condition 1, and denote the approximate solution to (16a) by $(\widehat{\Delta\lambda}, \widehat{\Delta x})$. Then for all sufficiently small μ, we have*

$$(\Delta\lambda, \Delta x, \Delta s) = O(\mu) \tag{31}$$

and

$$(\Delta\lambda - \widehat{\Delta\lambda}, \Delta x_B - \widehat{\Delta x}_B) = O(\mathbf{u}), \qquad \Delta x_N - \widehat{\Delta x}_N = O(\mu\mathbf{u}). \tag{32}$$

Proof. We prove (31) by appealing to (5). By partitioning A into B and N according to (10), and partitioning the diagonal matrices S and X accordingly, we see that the matrix in (5) is a permutation of

$$\begin{bmatrix} 0 & B & N & 0 & 0 \\ B^T & 0 & 0 & I & 0 \\ N^T & 0 & 0 & 0 & I \\ 0 & S_B & 0 & X_B & 0 \\ 0 & 0 & S_N & 0 & X_N \end{bmatrix}. \tag{33}$$

Because of (11), the diagonal elements in X_B and S_N are $\Omega(1)$, while the matrices S_B and X_N are $O(\mu)$. In addition, B is square and well conditioned, so the matrix (33) is an $O(\mu)$ perturbation of a uniformly nonsingular matrix. From (5), we then have

$$(\Delta\lambda, \Delta x, \Delta s) = O(\|r_b\| + \|r_c\| + \|XSe - \sigma\mu e\|),$$

so the result (31) follows from (11) and (12).

To derive the relative error estimate (32), consider the system (16a). By permuting the matrix in accord with the $\mathcal{B} \cup \mathcal{N}$ partition, we can rewrite (16a) as follows:

$$\begin{bmatrix} 0 & B & N \\ B^T & -X_B^{-1}S_B & \\ N^T & & -X_N^{-1}S_N \end{bmatrix} \begin{bmatrix} \Delta\lambda \\ \Delta x_B \\ \Delta x_N \end{bmatrix} = \begin{bmatrix} -r_b \\ -(r_c)_B + s_B - \sigma\mu X_B^{-1} e \\ -(r_c)_N + s_N - \sigma\mu X_N^{-1} e \end{bmatrix}. \tag{34}$$

From (11), we have for sufficiently small μ that the diagonals in $X_B^{-1}S_B$ are $\Omega(\mu)$ while the diagonals of $X_N^{-1}S_N$ are $\Omega(\mu^{-1})$, so this coefficient matrix is canonical.

By defining

$$M_B = \begin{bmatrix} 0 & B \\ B^T & -X_B^{-1} S_B \end{bmatrix}, \qquad M_N = \begin{bmatrix} N \\ 0 \end{bmatrix},$$

$$\Lambda = -X_N^{-1} S_N, \qquad z_N = \Delta x_N, \qquad z_B = \begin{bmatrix} \Delta\lambda \\ \Delta x_B \end{bmatrix},$$

$$d_B = \begin{bmatrix} -r_b \\ -(r_c)_B + s_B - \sigma\mu X_B^{-1} e \end{bmatrix}, \qquad d_N = -(r_c)_N + s_N - \sigma\mu X_N^{-1} e,$$

we can restate the system as

$$\begin{bmatrix} M_B & M_N \\ M_N^T & \Lambda \end{bmatrix} \begin{bmatrix} z_B \\ z_N \end{bmatrix} = \begin{bmatrix} d_B \\ d_N \end{bmatrix}. \tag{35}$$

From our assumption on B, we have $M_B = O(1)$ and $M_B^{-1} = O(1)$.

Because of Condition 1, the computed solution $\hat{z}$ of (35) satisfies

$$\left(\begin{bmatrix} M_B & M_N \\ M_N^T & \Lambda \end{bmatrix} + \Delta \right) \begin{bmatrix} \hat{z}_B \\ \hat{z}_N \end{bmatrix} = \begin{bmatrix} \hat{d}_B \\ \hat{d}_N \end{bmatrix}, \tag{36}$$

where $\hat{d} - d = O(\mathbf{u})$ and the canonical error matrix Δ satisfies

$$|\Delta| \le O(\mathbf{u}) + \begin{bmatrix} |M_B| & |M_N| \\ |M_N|^T & |\Lambda| \end{bmatrix} O(\mathbf{u}) = O(\mathbf{u}) + \begin{bmatrix} 0 & 0 \\ 0 & |\Lambda| \end{bmatrix} O(\mathbf{u}).$$

By combining this estimate with (35) and (36), we obtain

$$\left(\begin{bmatrix} M_B & M_N \\ M_N^T & \Lambda \end{bmatrix} + \Delta \right) \begin{bmatrix} \hat{z}_B - z_B \\ \hat{z}_N - z_N \end{bmatrix} = -\Delta \begin{bmatrix} z_B \\ z_N \end{bmatrix} + \begin{bmatrix} \hat{d}_B - d_B \\ \hat{d}_N - d_N \end{bmatrix}. \tag{37}$$

Since $z = O(\mu)$ from (31), we have

$$\left| \Delta \begin{bmatrix} z_B \\ z_N \end{bmatrix} \right| \le \left(O(\mathbf{u}) + \begin{bmatrix} 0 & 0 \\ 0 & |\Lambda| \end{bmatrix} O(\mathbf{u}) \right) \begin{bmatrix} O(\mu) \\ O(\mu) \end{bmatrix} \le \begin{bmatrix} O(\mu\mathbf{u}) \\ O(\mathbf{u}) \end{bmatrix}, \tag{38}$$

so when we add the effect of $\hat{d} - d$, we find that the right-hand side of (37) is $O(\mathbf{u})$. For the coefficient matrix in (37) we use Lemma 3.2 with

$$\begin{aligned} H_{11} &= M_B + O(\mathbf{u}), \\ H_{12} &= M_N + O(\mathbf{u}) = O(1), \\ H_{21} &= M_N^T + O(\mathbf{u}) = O(1), \\ H_{22} &= O(\mathbf{u}) + \Lambda(I + O(\mathbf{u})) = \Lambda(I + O(\mathbf{u})). \end{aligned}$$

Lemma 3.2 yields the following estimates:

$$\begin{aligned}
(H^{-1})_{22} &= \left(H_{22} - H_{21}H_{11}^{-1}H_{12}\right)^{-1} = \Lambda^{-1}(I + O(\mathbf{u}+\mu)) = O(\mu), \\
(H^{-1})_{12} &= O(\mu), \qquad (H^{-1})_{21} = O(\mu), \\
(H^{-1})_{11} &= M_B^{-1} + O(\mu + \mathbf{u}).
\end{aligned}$$

By combining these observations with (38), we obtain

$$\begin{bmatrix} \hat{z}_B - z_B \\ \hat{z}_N - z_N \end{bmatrix} = \begin{bmatrix} (H^{-1})_{11} & (H^{-1})_{12} \\ (H^{-1})_{21} & (H^{-1})_{22} \end{bmatrix} O(\mathbf{u}) = \begin{bmatrix} O(\mathbf{u}) \\ O(\mu\mathbf{u}) \end{bmatrix},$$

giving (32). ■

Next, we examine the accuracy of $\widehat{\Delta s}$, which is calculated by substituting $\widehat{\Delta\lambda}$ and $\widehat{\Delta x}$ into (16b).

Theorem 5.2 *Suppose that the assumptions of Theorem 5.1 are satisfied and that $\widehat{\Delta s}$ is evaluated in floating-point arithmetic from the formula (16b), with $\widehat{\Delta x}$ replacing Δx. We then have*

$$\Delta s_B - \widehat{\Delta s}_B = O(\mu\mathbf{u}), \tag{39a}$$

$$\Delta s_N - \widehat{\Delta s}_N = O(\mathbf{u}). \tag{39b}$$

Proof. Standard roundoff error analysis applied to (16b) shows that

$$\widehat{\Delta s} = -s + \sigma\mu X^{-1}e - X^{-1}S\widehat{\Delta x} + \left[|s| + \sigma\mu|X^{-1}|e + |X^{-1}S||\widehat{\Delta x}|\right] O(\mathbf{u}). \tag{40}$$

By differencing (16b) and (40), we obtain

$$|\Delta s - \widehat{\Delta s}| \le |X^{-1}S||\Delta x - \widehat{\Delta x}| + \left[|s| + \sigma\mu|X^{-1}|e + |X^{-1}S||\widehat{\Delta x}|\right] O(\mathbf{u}). \tag{41}$$

If $i \in \mathcal{B}$, we have from (11) that

$$|X_B^{-1}S_B| = O(\mu), \qquad |s_B| = O(\mu), \qquad \sigma\mu|X_B^{-1}|e = O(\mu).$$

By combining these estimates with (32) and (41), we obtain the desired result (39a). For $i \in \mathcal{N}$, we have from (11) again that

$$|X_N^{-1}S_N| = O(\mu^{-1}), \qquad |s_N| = O(1), \qquad \sigma\mu|X_N^{-1}|e = O(1),$$

while from (31) and (32) we have

$$\Delta x_N - \widehat{\Delta x}_N = O(\mu\mathbf{u}), \qquad \widehat{\Delta x}_N = O(\mu).$$

By substituting in (41), we obtain (39b). ■

The last two results show that the requirements of Theorem 4.1 are satisfied, so that the algorithm can make significant progress along these search directions. We summarize the combination of Theorems 4.1, 5.1, and 5.2 as a corollary.

Corollary 5.3 *Suppose that Assumption 1 holds and that the problem is nondegenerate, that is $|\mathcal{B}| = m$ with $\kappa(B)$ moderate. Suppose that the procedure for solving (16) satisfies Condition 1. If the approximate step is computed with $\sigma \in [0, 1/2]$, then for all sufficiently small μ, the formulae (19), (20), and (21) are satisfied.*

6 The Bunch-Kaufman Factorization

We show in this section that a procedure for solving (16a) based on the Bunch-Kaufman factorization satisfies Condition 1, so that the conclusion of Corollary 5.3 applies. Since much of the analysis of this section can be reused in the analysis of the Bunch-Parlett and sparse Bunch-Parlett algorithms, we give the details here and refer to them in later sections.

It is sufficient to describe just the first stage of the procedure. Later stages apply the same technique recursively to the remaining submatrix.

The pivot selection procedure for Bunch-Kaufman [1] is as follows.

```
Choose δ ∈ (0,1); find r such that χ_1 = |T_{r1}|;
if  χ_1 > 0
      if  |T_{11}| ≥ δχ_1
            1 × 1 pivot, P_1 = I
      else
            find χ_r;
            if  χ_r|T_{11}| ≥ δχ_1^2
                  1 × 1 pivot, P_1 = I
            elseif  |T_{rr}| ≥ δχ_r
                  1 × 1 pivot; choose P_1 so that (P_1 T P_1^T)_{11} = T_{rr}
            else
                  2 × 2 pivot; choose P_1 so that (P_1 T P_1^T)_{21} = T_{r1}
            end
      end
end.
```

If we denote the 1×1 or 2×2 pivot block by E and write

$$P_1 T P_1^T = \begin{bmatrix} E & C^T \\ C & \hat{T} \end{bmatrix}, \tag{42}$$

the first step of the factorization yields

$$P_1 T P_1^T = \begin{bmatrix} I & \\ CE^{-1} & I \end{bmatrix} \begin{bmatrix} E & \\ & \bar{T} \end{bmatrix} \begin{bmatrix} I & E^{-1}C^T \\ & I \end{bmatrix}, \tag{43}$$

where $\bar{T} = \hat{T} - CE^{-1}C$. The algorithm continues by applying this procedure to $\bar{T}$. Note that the χ_i are generally changed by each stage of the factorization. The submatrix CE^{-1} contains the subdiagonals in the first one or two columns of the L factor.

Bunch and Kaufman [1] show that for the particular choice $\delta = (1 + \sqrt{17})/8$, we have

$$\max_{i,j} |\bar{T}_{ij}| \leq (2.57) \max_{i,j} |T_{ij}|, \tag{44}$$

so there is a modest bound on element growth during each stage of the factorization.

When applied to canonical matrices, the Bunch-Kaufman procedure selects pivots of specific types and produces a reduced submatrix that is also canonical. We state these results in the following two theorems, whose proofs are tedious and are relegated to the Appendix.

Theorem 6.1 *Let one step of the Bunch-Kaufman factorization be applied to a canonical matrix that is not degenerate. Then*

(a) The pivot block E will be either

(i) a 1×1 block, chosen from among the diagonal elements of Λ; or

(ii) a 2×2 block, in which the off-diagonal element E_{12} is one of the elements of B;

(b) The matrix remaining after the elimination is canonical, and the absolute change in the elements of Λ is at most $O(1)$;

(c) Using the notation from (42), we have that $|C| = O(1)$, while

(i) $|E| = O(\mu^{-1})$ and $|E^{-1}| = O(\mu)$ if E is a 1×1 pivot; and

(ii) $|E| = O(1)$ and $|E^{-1}| = O(1)$ if E is a 2×2 pivot.

Theorem 6.2 *Let one step of the Bunch-Kaufman factorization be applied to a degenerate canonical matrix. Then*

(a) The pivot block E will be either

(i) a 1×1 block, chosen from any of the diagonals (large or small); or

(ii) a 2×2 block, in which all the elements are $O(\mu + \mathbf{u})$;

(b) The matrix remaining after the elimination is canonical (not necessarily degenerate), and the absolute change to the remaining matrix is $O(\mu + \mathbf{u})$.

Because of Assumption 1, our initial matrix in (16a) is canonical. Barring pathological growth in the remaining submatrices, one of Theorems 6.1 and 6.2 applies at every stage of the Bunch-Kaufman factorization.

If B is *square* in the original matrix (corresponding to a nondegenerate linear program), then the remaining matrices encountered at every stage of the factorization are not degenerate. After a 1×1 pivot, the dimensions of B are unchanged, while a 2×2 pivot shrinks B by exactly one row and column, so it remains square. When a pivot causes B to disappear altogether, the reduced matrix has the form $\Lambda + O(\mu + \mathbf{u})$. It follows that in the case of square B, Theorem 6.1 is sufficient to analyze the entire factorization. The following result gives the backward error analysis for the factorization in this case.

Corollary 6.3 *Let the Bunch-Kaufman factorization be applied to a canonical matrix T in which B is square. Then, for all sufficiently small μ, we obtain computed factors $\hat{L}$ and $\hat{D}$ such that*

$$\hat{L}\hat{D}\hat{L}^T = PTP^T + P\bar{\Delta}P^T, \tag{45}$$

where $\bar{\Delta}$ is a canonical error matrix associated with T.

Proof. We prove the result by an induction argument on the dimension $\bar{n} = m + n$ of the matrix T. The induction is made slightly more complex than usual by the form of the canonical matrix, notably, the presence of the square matrix B of dimension $m \leq n$.

For $\bar{n} = 1$, we must have $m = 0$ and so trivially $P = 1$, $\hat{L} = 1$, $\hat{D} = T_{11}$. Therefore (45) holds with $\bar{\Delta} = 0$.

For $\bar{n} = 2$, we have two cases $m = 0$ and $m = 1$. For $m = 0$, there are two elements of magnitude $\Omega(\mu^{-1})$ on the diagonal, while the off-diagonals are $O(\mu + \mathbf{u})$. Hence, a 1×1 pivot is chosen. If there is no pivoting, the first step of elimination yields

$$\begin{aligned}
\hat{L}_{21} &= T_{21}/T_{11} + |T_{21}/T_{11}|O(\mathbf{u}), \\
\hat{D}_{11} &= T_{11}, \\
\hat{D}_{22} &= T_{22} - T_{21}^2/T_{11} + (|T_{22}| + |T_{21}^2/T_{11}|)O(\mathbf{u}).
\end{aligned}$$

Since $\hat{L}$ has unit diagonals, we obtain by expanding the factors that

$$\hat{L}\hat{D}\hat{L}^T = T + \begin{bmatrix} 0 & |T_{21}|O(\mathbf{u}) \\ |T_{21}|O(\mathbf{u}) & |T_{21}^2/T_{11}|O(\mathbf{u}) + |T_{22}|O(\mathbf{u}) \end{bmatrix} = T + \bar{\Delta},$$

where

$$|\bar{\Delta}| \leq |T|O(\mathbf{u}) + O(\mathbf{u}),$$

so $\bar{\Delta}$ is a canonical error matrix associated with T. The same logic applies if pivoting occurs.

In the remaining case $m = 1$, the pivot is 2×2, we have $\hat{L} = I$, $P = I$, and $\hat{D} = T$, and (45) holds trivially with $\bar{\Delta} = 0$.

We now examine a canonical matrix of dimension $\bar{n} > 2$ in which B is square, and examine the first stage of the factorization. Because the matrix is canonical and nondegenerate, Theorem 6.1 applies. For some permutation matrix P_1, we have from (42) and (43) that the first stage yields partial factors $\hat{L}_1$ and $\hat{D}_1$, where

$$\hat{L}_1 = \begin{bmatrix} I & 0 \\ CE^{-1} + \Delta_L & I \end{bmatrix}, \qquad \hat{D}_1 = \begin{bmatrix} E & 0 \\ 0 & \bar{T} + \Delta_D \end{bmatrix}, \tag{46}$$

where

$$|\Delta_L| \leq |C||E^{-1}|O(\mathbf{u}), \qquad |\Delta_D| \leq |\hat{T}|O(\mathbf{u}) + |C||E^{-1}||C|^T O(\mathbf{u}) = |\hat{T}|O(\mathbf{u}) + O(\mathbf{u}).$$

Note that Δ_D is a canonical error matrix corresponding to $\hat{T}$. By the proof of Theorem 6.1, the $(2,2)$ submatrix of $\hat{D}_1$ is canonical, so we use the inductive hypothesis to deduce that the $\hat{L}$, $\hat{D}$ factors of this submatrix satisfy

$$\hat{L}_2\hat{D}_2\hat{L}_2^T = P_2(\bar{T} + \Delta_D)P_2^T + P_2\bar{\Delta}_2P_2^T \tag{47}$$

for some permutation matrix P_2 and some canonical error matrix $\bar{\Delta}_2$ corresponding to $(\bar{T} + \Delta_D)$. We compose the overall factors of T as follows:

$$\hat{L} = \begin{bmatrix} I & 0 \\ P_2(CE^{-1} + \Delta_L) & \hat{L}_2 \end{bmatrix}, \qquad \hat{D} = \begin{bmatrix} E & 0 \\ 0 & \hat{D}_2 \end{bmatrix}, \qquad P = \begin{bmatrix} I & 0 \\ 0 & P_2 \end{bmatrix} P_1.$$

Now,

$$\hat{L}\hat{D}\hat{L}^T = \begin{bmatrix} E & (C+\Delta_2)^T P_2^T \\ P_2(C+\Delta_2) & \hat{L}_2\hat{D}_2\hat{L}_2^T + P_2CE^{-1}C^TP_2^T + P_2\Delta_1P_2^T \end{bmatrix}, \tag{48}$$

where

$$\Delta_1 = \Delta_L C^T + C\Delta_L^T + \Delta_L E \Delta_L^T, \qquad \Delta_2 = \Delta_L E,$$

and so

$$|\Delta_1| \le |C||E^{-1}||C|^T O(\mathbf{u}) = O(\mathbf{u}), \qquad |\Delta_2| \le |C||E^{-1}||E| O(\mathbf{u}) = O(\mathbf{u}).$$

By substituting (47) and (46) into (48), we obtain

$$\begin{aligned} \hat{L}\hat{D}\hat{L}^T &= \begin{bmatrix} E & (C+\Delta_2)^T P_2^T \\ P_2(C+\Delta_2) & P_2\left[\bar{T} + \Delta_D + CE^{-1}C^T + \Delta_1 + \bar{\Delta}_2\right] P_2^T \end{bmatrix} \\ &= \begin{bmatrix} E & (C+\Delta_2)^T P_2^T \\ P_2(C+\Delta_2) & P_2\left[\hat{T} + \Delta_D + \Delta_1 + \bar{\Delta}_2\right] P_2^T \end{bmatrix} \\ &= PTP^T + P\bar{\Delta}P^T, \end{aligned}$$

where

$$\bar{\Delta} = P_1^T \begin{bmatrix} 0 & \Delta_2^T \\ \Delta_2 & \Delta_D + \Delta_1 + \bar{\Delta}_2 \end{bmatrix} P_1.$$

Since $|\Delta_1| = O(\mathbf{u})$, $|\Delta_2| = O(\mathbf{u})$, and Δ_D and $\bar{\Delta}_2$ are canonical error matrices corresponding to $\hat{T}$, we have

$$\begin{aligned} |\bar{\Delta}| &\le P_1^T \begin{bmatrix} 0 & |\Delta_2|^T \\ |\Delta_2| & |\Delta_D| + |\Delta_1| + |\bar{\Delta}_2| \end{bmatrix} P_1 \\ &\le O(\mathbf{u}) + P_1^T \begin{bmatrix} 0 & 0 \\ 0 & |\hat{T}| \end{bmatrix} P_1 O(\mathbf{u}) \\ &\le O(\mathbf{u}) + |T| O(\mathbf{u}). \end{aligned}$$

Hence, $\bar{\Delta}$ is a canonical error matrix corresponding to T.

We complete the proof by noting that Theorem 6.1 can be applied to the remaining matrix, because it is also canonical and nondegenerate. ■

Given the system $Tz = d$ and the data P, $\hat{L}$, and $\hat{D}$ from the factorization, the computed solution $\hat{z}$ is found by performing two vector permutations with P, triangular substitutions with $\hat{L}$ and $\hat{L}^T$, and a blockwise inversion of $\hat{D}$. The 2×2 diagonal blocks in $\hat{D}$ can be handled by the Gaussian elimination procedure outlined in the following technical lemma, which is proved in Appendix A.3. It is easy to show that the elements of the pivot block E satisfy the condition (49).

Lemma 6.4 *Consider the 2×2 linear system $Ey = g$ in which E is symmetric with*

$$|E_{11}| \le \delta |E_{12}|, \qquad |E_{11}||E_{22}| \le \delta^2 |E_{12}|^2, \tag{49}$$

for some $\delta \in (0,1)$. Then if we compute the solution by applying Gaussian elimination to the permuted system

$$\begin{bmatrix} E_{12} & E_{22} \\ E_{11} & E_{12} \end{bmatrix} \begin{bmatrix} y_1 \\ y_2 \end{bmatrix} = \begin{bmatrix} g_2 \\ g_1 \end{bmatrix}, \tag{50}$$

then the computed solution $\hat{y}$ satisfies

$$(E + \Delta_E)\hat{y} = g,$$

where

$$|\Delta_E| \le |E| O(\mathbf{u}). \tag{51}$$

The additional error that is introduced during recovery of the solution with the computed factors $\hat{L}$, $\hat{D}$, and $\hat{L}^T$ is quantified in the next result.

Lemma 6.5 *Suppose the assumptions and notation of Corollary 6.3 hold. Then the computed solution $\hat{z}$ to the system $\hat{L}\hat{D}\hat{L}^T z = Pd$ satisfies*

$$(\hat{L}\hat{D}\hat{L}^T + P\hat{\Delta}P^T)\hat{z} = Pd, \tag{52}$$

where $\hat{\Delta}$ is a canonical error matrix associated with T.

Proof. From standard results for triangular substitution, the computed solution of $\hat{L}z_a = Pd$ satisfies

$$(\hat{L} + \hat{\Delta}_{L1})\hat{z}_a = Pd, \qquad |\hat{\Delta}_{L1}| \le |\hat{L}| O(\mathbf{u}).$$

A similar result holds for triangular substitution with the transpose $\hat{L}^T$.

For solution of $\hat{D}z_b = \hat{z}_a$, we note that $\hat{D}$ is block-diagonal with 1×1 and 2×2 blocks. For the 2×2 pivot blocks that arise in the Bunch-Kaufman procedure, the assumptions of Lemma 6.4 hold, so the computed solution $\hat{y}$ of a 2×2 subsystem $Ey = g$ satisfies

$$(E + \Delta_E)\hat{y} = g, \qquad |\Delta_E| = |E| O(\mathbf{u}). \tag{53}$$

When E is a 1×1 block, the estimate (53) holds trivially. Hence, the computed solution $\hat{z}_b$ of $\hat{D}z_b = \hat{z}_a$ satisfies

$$(\hat{D} + \hat{\Delta}_D)\hat{z}_b = \hat{z}_a, \qquad |\hat{\Delta}_D| \le |\hat{D}| O(\mathbf{u}).$$

By combining the error expressions for the three component systems, we find that our computed solution $\hat{z}$ satisfies

$$(\hat{L} + \hat{\Delta}_{L1})(\hat{D} + \hat{\Delta}_D)(\hat{L} + \hat{\Delta}_{L2})^T \hat{z} = Pd.$$

Multiplying the matrix products, we find that (52) is satisfied with

$$P|\hat{\Delta}|P^T \le |\hat{L}||\hat{D}||\hat{L}|^T O(\mathbf{u}) + O(\mu\mathbf{u} + \mathbf{u}^2).$$

From our earlier discussions on the composition of $\hat{L}$ and $\hat{D}$, it is easy to see that the absolute matrix product $|\hat{L}||\hat{D}||\hat{L}|^T$ contains all $O(1)$ elements, except for the large diagonals, which occur in the same positions as in PTP^T. Hence $P\hat{\Delta}P^T$ is a canonical error matrix corresponding to PTP^T, and our proof is complete. ∎

We can now summarize the effects of roundoff error on the entire solution process for (16) in the following theorem.

Theorem 6.6 *Suppose T is a canonical matrix in which B is square. Then, for all sufficiently small μ, the Bunch-Kaufman factorization followed by the solution process outlined above satisfies Condition 1.*

Proof. As we noted immediately following Condition 1, the actual right-hand side may differ by terms of $O(\mathbf{u})$ from its "theoretical" value d. From (52), the computed solution $\hat{z}$ to $Tz = d$ satisfies

$$(\hat{L}\hat{D}\hat{L}^T + P\hat{\Delta}P^T)\hat{z} = \hat{d},$$

Substituting from (45), we obtain

$$(PTP^T + P\bar{\Delta}P^T + P\hat{\Delta}P^T)\hat{z} = P\hat{d},$$

so Condition 1 follows when we set $\Delta = \bar{\Delta} + \hat{\Delta}$. ∎

We have shown that in the case of a nondegenerate linear program, the procedure based on applying Bunch-Kaufman to (16a) leads to approximate steps $(\widehat{\Delta\lambda}, \widehat{\Delta x}, \widehat{\Delta s})$ that satisfy the conditions of Theorem 4.1. The estimate (20) implies that during the final iterations of a primal-dual algorithm, near-unit steps can be taken along these directions without leaving the nonnegative orthant. Moreover, if the centering parameter σ is small or zero, a large reduction in the duality gap μ can be expected. In the extreme case $\sigma = 0$ (the "affine-scaling" choice), linear convergence with a rate constant of $O(\mathbf{u})$ can be attained if the actual step length is close to $\hat{\alpha}^*$. Most practical algorithms choose the step length to be a fixed multiple — typically .95 or .9995 — of $\hat{\alpha}^*$, and indeed these methods often converge rapidly during their final stages. For algorithms that use a more theoretically justifiable definition of step length the story is not, unfortunately, this simple. In [21, Section 4], for instance, extra restrictions are applied to α to ensure that (12) and (14) continue to hold at the next iterate. These restrictions may result in α being much smaller than one. This case is analyzed in [21, Section 4], so we do not repeat it here.

7 The Bunch-Parlett Factorization

The Bunch-Parlett searches the entire remaining matrix for each pivot, not just one or two columns. The pivot selection procedure is as follows.

Choose $\delta \in (0,1)$, $\chi_{\text{off}} = |T_{rs}| = \max_{i \neq j} |T_{ij}|$, $\chi_{\text{diag}} = |T_{pp}| = \max_i |T_{ii}|$;
if $\chi_{\text{diag}} \geq \delta \chi_{\text{off}}$

$s = 1$ and choose P_1 so that $(P_1 T P_1^T)_{11} = T_{pp}$

else

$s = 2$ and choose P_1 so that $(P_1 T P_1^T)_{21} = T_{rs}$

end.

The elimination step is identical to Bunch-Kaufman, and the process of using the LDL^T factorization to solve the system $Tz = d$ is the same as in the preceding section. As in Bunch-Kaufman, the value $\delta = (1 + \sqrt{17})/8$ leads to the modest bound of 2.57 on element growth at each stage.

When applied to canonical matrices, the Bunch-Parlett factorization proceeds in three stages:

1. All the diagonal elements of Λ are selected as 1×1 pivots;
2. 2×2 pivots of the type described in Theorem 6.1(a) are chosen;
3. When no more 2×2 pivots like this are available and the remaining matrix contains only elements of size $O(\mu + \mathbf{u})$, a combination of small 1×1 and 2×2 pivots is used to complete the factorization process.

We prove this assertion in the following lemma.

Theorem 7.1 *Suppose that the Bunch-Parlett procedure is applied to a canonical matrix. Then the factorization proceeds according to the three-stage outline above. If the canonical matrix has B square and is nonvacuous, the factorization is completed by stages 1 and 2; stage 3 is vacuous.*

Proof. Assuming that Λ is not vacuous, we have at the pivot selection step that $\chi_{\text{off}} = O(1)$ and $\chi_{\text{diag}} = \Omega(\mu^{-1})$. The pivot element will therefore be one of the large diagonals corresponding to Λ. The remaining matrix is updated by subtracting $CE^{-1}C$, where clearly $C = O(1)$ and $E^{-1} = O(\mu)$. Hence, the remaining matrix retains canonical form.

We can apply this argument inductively until all the diagonals in Λ are exhausted. At the end of stage 1, the remaining matrix has the form

$$\begin{bmatrix} 0 & B \\ B^T & 0 \end{bmatrix} + O(\mu + \mathbf{u}). \tag{54}$$

Stage 2 now begins. If B is not vacuous, we have $\chi_{\text{off}} = O(1)$ and $\chi_{\text{diag}} = O(\mu + \mathbf{u})$. In fact, by the assumption $B = \Omega(1)$, we have $\chi_{\text{off}} = \Omega(1)$, and the element T_{rs} that achieves the maximum comes from B. The 2×2 block with off-diagonal element T_{rs} is selected as the pivot. After the elimination step, the size of B is reduced by one row and column. The proof of Theorem 6.1(b) can be applied again here to show that the remaining matrix is also canonical, so 2×2 pivots of this type will continue to be selected until B vanishes.

The number of steps in stage 2 is $\min(\text{rows}(B), \text{columns}(B))$. At the end of this stage, the remaining matrix is square with dimension $|\text{rows}(B) - \text{columns}(B)|$, and all its elements

have size $O(\mu + \mathbf{u})$. In stage 3, both 1×1 and 2×2 pivots may be used to factor this matrix. If B is square, the factorization is complete after stage 2. ∎

The other major results of Section 6 continue to hold when the Bunch-Parlett algorithm is used instead of Bunch-Kaufman; only trivial adjustments to the analysis in Section 6 and Appendix A.1 are necessary. We summarize the conclusions in the following theorem.

Theorem 7.2 *Suppose T is a canonical matrix in which B is square. Then, for all sufficiently small μ, the Bunch-Parlett factorization followed by the solution process outlined in Section 6 satisfies Condition 1.*

8 The Sparse Bunch-Parlett Factorization

Several authors (notably Fourer and Mehrotra [4]) have proposed a sparse variant of the Bunch-Parlett factorization that compromises between maintaining sparsity and limiting element growth in the remaining matrix. We outline the pivot selection procedure as described by [4], with a slight modification noted below.

For each index $i = 1, 2, \ldots, \bar{n}$ we define the *degree* n_i to be the number of off-diagonal nonzeros in row i. We also define an estimate of the joint nonzero content of rows i and j by

$$\hat{n}_{ij} = \min(n_i + n_j - 4, \bar{n} - 2).$$

A 2×2 pivot block

$$E = \begin{bmatrix} T_{ii} & T_{ij} \\ T_{ij} & T_{jj} \end{bmatrix} \tag{55}$$

is termed *oxo* if both of T_{ii} and T_{jj} are zero, *tile* if one of T_{ii} and T_{jj} is zero, and *full* if both of T_{ii} and T_{jj} are nonzero. We define a *cost* associated with using (55) as the pivot block in each of these three cases by

$$\begin{array}{ll} \textit{oxo:} & (n_i - 1)(n_j - 1), \\ \textit{tile:} & (n_i - 1)(\hat{n}_{ij} + 1) \quad \text{if } T_{ii} = 0, \qquad (n_j - 1)(\hat{n}_{ij} + 1) \quad \text{if } T_{jj} = 0, \\ \textit{full:} & \hat{n}_{ij}^2, \end{array}$$

The cost is an estimate of the fill-in associated with using (55) as the pivot block.

For prospective pivots, we define stability criteria in terms of the usual constant $\delta \in (0, 1)$ and the off-diagonal norms χ_i defined in (29). Any 1×1 pivot must satisfy

$$|T_{ii}^{-1}|\chi_i \leq 2/\delta, \tag{56}$$

while a 2×2 pivot (55) must have

$$\left| \begin{bmatrix} T_{ii} & T_{ij} \\ T_{ij} & T_{jj} \end{bmatrix}^{-1} \right| \begin{bmatrix} \chi_i \\ \chi_j \end{bmatrix} \leq \begin{bmatrix} 1/\delta \\ 1/\delta \end{bmatrix}. \tag{57}$$

The pivot selection procedure is as follows.

```
for  r = 1, 2, ...
      for  i with n_i = r
            consider T_ii with degree r;
            if  any of these elements satisfy (56)
                  accept as a 1 × 1 pivot and exit;
            else  label it as unstable;
      end

      for  unstable pivots T_ii from the previous loop
            consider 2 × 2 pivots involving T_ii, with costs at most
                  (r − 1)^2, (r − 1)(2r − 3), and (2r − 4)^2
                  for oxo, tile, and full pivots, respectively;
            if  any of these blocks satisfy (57)
                  accept as a 2 × 2 pivot and exit;
      end
end.
```

The pivot selection pattern for the sparse Bunch-Parlett algorithm is essentially the same as for the Bunch-Kaufman algorithm, as described in Theorems 6.1 and 6.2. We prove this result in the appendix, since the analysis differs a little from the Bunch-Kaufman case.

Theorem 8.1 *The results of Theorems 6.1 and 6.2 hold when the sparse Bunch-Parlett factorization is used in place of the Bunch-Kaufman procedure.*

To obtain this result, we modified the acceptance condition (56) for 1×1 pivots. In the description of [4], the right-hand side is $1/\delta$ rather than $2/\delta$. With the original choice, the sparse Bunch-Parlett algorithm applied to a degenerate canonical matrix could allow another type of pivot: a 2×2 pivot in which one diagonal is from Λ and the other has size $O(\mu + \mathbf{u})$. A pivot of this type is poorly conditioned and will generally lead to instability during the blockwise inversion of $\hat{D}$.

The other major results of Section 6 also continue to hold when the sparse Bunch-Parlett algorithm is used in place of Bunch-Kaufman. We summarize the conclusions in the following theorem.

Theorem 8.2 *Suppose T is a canonical matrix in which B is square. Then, for all sufficiently small μ, the sparse Bunch-Parlett factorization followed by the solution process outlined in Section 6 satisfies Condition 1.*

9 The Degenerate Case

When the linear program (1), (2) is degenerate — $|\mathcal{B}| \neq m$ — the three factorization procedures can no longer run to completion with just the two kinds of pivots described in Theorem

6.1. The nonsquare shape of B in the matrix (34) means that pivots of size $O(\mu+\mathbf{u})$ — either 1×1 or 2×2 — are used at some point in the factorization process. The factorizations fail only if these pivots are exactly zero, which happens often on small problems but not otherwise. The more common outcome is that the interior-point algorithm makes only slow or erratic progress after μ has achieved a certain (small) value. In this section we sketch the reasons for this outcome.

In all the factorizations above, the large diagonal elements in $X_N^{-1}S_N$ are used as 1×1 pivots. Even though these pivots are not necessarily used before any others (except in the Bunch-Parlett algorithm), the factorizations behave as if they were solving the system (16) in the equivalent, partitioned form

$$\begin{bmatrix} NX_NS_N^{-1}N^T & B \\ B^T & -X_B^{-1}S_B \end{bmatrix}\begin{bmatrix} \Delta\lambda \\ \Delta x_B \end{bmatrix} = \begin{bmatrix} -r_b + NS_N^{-1}X_N[-(r_c)_N + s_N - \sigma\mu X_N^{-1}e] \\ -(r_c)_B + s_B - \sigma\mu X_B^{-1}e \end{bmatrix}, \tag{58}$$

$$\Delta x_N = X_N^{-1}S_N\left[(r_c)_N - s_N + \sigma\mu X_N^{-1}e + N^T\Delta\lambda\right]. \tag{59}$$

The coefficient matrix in (58) is an $O(\mu)$ perturbation of the matrix

$$\begin{bmatrix} 0 & B \\ B^T & 0 \end{bmatrix}. \tag{60}$$

Since B is well conditioned by Definition 2, the matrix in (60) has $2\min(|\mathcal{B}|, m)$ nonzero singular values of magnitude $\Omega(1)$. In the nondegenerate case, (60) is well conditioned. Otherwise, it has $|m-|\mathcal{B}||$ zero singular values. When $|\mathcal{B}|<m$, the null space of (60) is spanned by

$$\begin{bmatrix} \bar{Z} \\ 0 \end{bmatrix}, \tag{61}$$

where $\bar{Z}$ is an $m\times(m-|\mathcal{B}|)$ matrix of full rank such that $B^T\bar{Z}=0$. When $|\mathcal{B}|>m$, the null space of (60) is spanned by the matrix

$$\begin{bmatrix} 0 \\ \hat{Z} \end{bmatrix}, \tag{62}$$

where $\hat{Z}$ spans the null space of B. For small μ, these null spaces are not altered much by the perturbation of size $O(\mu)$ that is present in the matrix (58), because the nonzero singular values of (60) are well separated from zero. Perturbations in the solution of (58) due to roundoff will occur mainly in the space of small singular values. Hence, when $|\mathcal{B}|<m$, the perturbations occur mostly in the range space of the matrix (61), that is, in the components of $\Delta\lambda$. Similarly, when $|\mathcal{B}|>m$, the perturbations occur in the range space of the matrix (61), that is, in the components of Δx_B.

The main source of difficulty is inaccuracy in the computed residual vectors r_b and r_c which, as mentioned above, contain errors of $O(\mathbf{u})$. In the case $|\mathcal{B}|>m$, these perturbations are magnified by the inverses of the small singular values, usually leading to errors of size

about $O(\mathbf{u}/\mu)$ in the components of Δx_B. The large relative errors in Δx_B induce large relative errors in Δs_B through the formula (16b). The step length to the boundary α^* may therefore be sharply curtailed because of the nonnegativity requirements (17a). In the case $|\mathcal{B}| < m$, the large relative errors in $\Delta\lambda$ induce errors in Δx_N through the formula (59), while in turn induce large relative errors in Δs_N through (16b). The step length may again be curtailed as a result.

Errors from sources other than the vector r are less significant.

If we have a strictly feasible starting point (see (13)), then we can simply set $r = 0$ throughout the algorithm. In this case, we can fix r at zero in the computations and avoid the problem above. It is usually not easy to find such a starting point, however, so some thought should be given to other ways of dealing with the problem.

One option is to simply terminate the algorithm when it stalls, declaring success if both μ and r are small. This option works well for most purposes, since stalling usually occurs only after μ is reduced to $O(\mathbf{u})$, by which time the problem has usually converged to acceptable accuracy. Fourer and Mehrotra [4] report that the convergence criteria are usually satisfied before the ill effects of roundoff are seen. Our testing in Section 10 allows a similar conclusion.

A second option is to switch to a termination procedure when the interior-point algorithm stalls. A finite termination procedure (see, for example, Ye [23]) or crossover to the simplex method (Meggido [12]) could be activated.

A third option is simply to fix r at zero in the computations once it has reached the $O(\mathbf{u})$ level, because at this stage our current point is feasible to within the limits of floating-point arithmetic. By doing so, we are effectively introducing a perturbation into the problem to freeze the infeasibility at its current level. This perturbation has an interesting effect: It moves the solution to a particular vertex of the previously optimal face, changing the $\mathcal{B} \cup \mathcal{N}$ partition appropriately. If we continue to run the interior-point algorithm to higher accuracy, it eventually converges to this vertex, but only after going through many more iterates (and taking some sharp turns in the process). The result of this process is similar to what we would achieve with a crossover to simplex, but the computational cost would generally be much higher.

10 Computational Experiments

We report here on some computational experiments that demonstrate the effects described above. Our testbed algorithm is the infeasible-interior-point path-following algorithm described in Wright [20]. In exact arithmetic, this algorithm achieves superlinear convergence because it eventually always takes affine-scaling steps ($\sigma = 0$ in (5)) with step length α approaching 1. This algorithm performs well on practical problems, but is not as fast as codes that use the Mehrotra predictor-corrector heuristic, for which no solid convergence theory exists, except in the nondegenerate case. The asymptotic behavior in finite precision is quite similar for the two algorithms.

To show that the finite precision effects are not confined to "nice" problems, we generate problems with fairly wide variations in the components of A, x_B^*, and s_N^*. The matrix A is

dense and random, with elements defined by

$$\begin{aligned} A_{1j} &= \tau 10^{6\tau-3}, \qquad j = 1, \ldots, n, \\ A_{ij} &= (\tau - .5)10^{6\tau-3}, \qquad i = 2, \ldots, m, \quad j = 1, \ldots, n, \end{aligned}$$

where every instance of τ is selected from a uniform distribution on the interval $[0, 1]$. (We choose all the elements in the first row of A to be positive to ensure that the feasible region is bounded.) We control the size of the index sets $\mathcal{B}$ and $\mathcal{N}$ (to control the amount of degeneracy) and set

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

We choose a particular solution (λ^*, x^*, s^*) by setting

$$\begin{gathered} \lambda^* = e, \qquad s^*_B = 0, \qquad x^*_N = 0, \\ s^*_i = 10^{4\tau-2}, \;\; i \in \mathcal{N}, \qquad x^*_i = 10^{3\tau-1}, \;\; i \in \mathcal{B}, \end{gathered}$$

where each τ is as before. The vectors b and c are determined by the choices of A and (λ^*, x^*, s^*).

The LAPACK Bunch-Kaufman factorization routines **dsytrf** and **dsytrs** are used to solve (16a). These routines (and the rest of our code) use double-precision arithmetic, giving $\mathbf{u} \approx 10^{-14}$ on the SPARC-5 on which these results were obtained.

We report on problems with $m = 6$, $n = 12$. (In problems smaller than this, exactly zero pivots often occur in degenerate cases, leading to breakdown.) Termination occurs when $\mu \leq 10^{-30}$ — an artificially stringent criterion, chosen to give us a clear look at asymptotic effects.

The first result is for a nondegenerate problem, for which $|\mathcal{B}| = m = 6$. Table 1 shows the sizes of μ and $\|r\|$ on each iterate. For the reasons that we outlined immediately following Condition 1, $\|r\|$ stabilizes at a magnitude of $O(\mathbf{u})$. The duality gap μ does not converge subquadratically (as it would in exact arithmetic) but rather exhibits extremely fast linear convergence, with a rate constant of about 10^{-10}. This is exactly the effect predicted by formula (21) for the affine-scaling steps that are taken on the last four iterations.

To see that the pivots have the properties predicted by Theorem 6.1, we examine the matrix D from the Bunch-Kaufman factorization. Table 2 shows D at iteration 17, when $\mu \approx 10^{-7}$. As expected, there are six 1×1 pivots of magnitude $\Omega(\mu^{-1})$, and six 2×2 pivots in which the diagonals are tiny and the off-diagonals are $\Omega(1)$. The same structure is present in D at every iteration after iteration 15.

Our second example is for a dual degenerate problem with $|\mathcal{B}| = 6 > m$. As can be seen from Table 3, the algorithm achieves fairly high accuracy after about 20 iterations, but no further improvement can be made after that point. The behavior is consistent with the discussion of Section 9. It suggests that the results of Section 6 are "tight," in that we cannot prove that "useful" search directions are obtained for arbitrarily small μ.

Examination of the D factor for the second example (Table 4) shows that the pivot pattern is in line with the predictions of Theorems 6.1 and 6.2. Together, these results imply

Table 1: Nondegenerate problem: $m = 6, n = 12$

k	$\log_{10} \mu_k$	$\log_{10} \|r^k\|_1$	Affine Step?
1	5.4	3.1	
2	4.7	2.3	
3	4.3	1.6	
4	3.8	0.8	
5	3.1	-12.0	
$\vdots$	$\vdots$	$\vdots$	
15	-3.2	-14.0	
16	-4.6	-13.7	$*$
17	-7.2	-14.4	$*$
18	-12.3	-14.1	$*$
19	-22.1	-13.8	$*$
20	-33.3	-14.2	termination

Table 2: The D factor at iteration 17 of the nondegenerate test problem ($*$ = magnitude less than 10^{-6})

Row/Column	Pivot Block	
1,2	$*$	.94(1)
	.94(1)	$*$
3,4	$*$	-.91(2)
	-.91(2)	$*$
5	.26(7)	
6	.30(11)	
7	.33(10)	
8	.47(7)	
9,10	-.30(-5)	.71(2)
	.71(2)	$*$
11,12	-.27(-3)	-.15(2)
	-.15(2)	$*$
13,14	$*$	-.31(0)
	-.31(0)	-.49(-5)
15,16	$*$	.16(0)
	.16(0)	$*$
17	.27(4)	
18	.32(6)	

Table 3: Dual degenerate problem: $m = 6, n = 12, |\mathcal{B}| = 8$

k	$\log_{10} \mu_k$	$\log_{10} \|r^k\|_1$	Affine Step?
1	5.4	3.1	
$\vdots$	$\vdots$	$\vdots$	
19	-6.0	-13.8	
20	-9.8	-14.1	*
21	-13.6	-14.2	*
22	-14.8	-13.8	*
23	-15.4	-13.2	*
$\vdots$	$\vdots$	$\vdots$	
99	-17.5	-13.5	
100	-17.5	-13.4	
$\vdots$	$\vdots$	$\vdots$	

that there are exactly $\min(m, |\mathcal{B}|)$ of the stable 2×2 pivots with an off-diagonal from B, and $|\mathcal{N}| = n - |\mathcal{B}|$ of the large 1×1 pivots. Together, these stable pivots account for

$$2 \min(m, |\mathcal{B}|) + |\mathcal{N}| = n + m - |m - |\mathcal{B}|| \tag{63}$$

stages of the factorization, so unstable pivots are used on the remaining submatrix whose dimension is $|m - |\mathcal{B}||$. In Table 4, we see that the last two 1×1 pivots are unstable, as expected. As we described in the first part of Section 9, the errors in $\widehat{\Delta x}_B$ and $\widehat{\Delta s}_B$ are preventing further progress. On iteration 100, the computed affine step has $\|\widehat{\Delta x}_B\|_\infty = .17(6)$, while its exact counterpart would have $\|\Delta x_B\|_\infty = O(\mu)$. By comparing components of $\widehat{\Delta s}_B$ with s_B, we find that the step to the boundary is sharply curtailed by the restriction $s_B + \alpha \widehat{\Delta s}_B \geq 0$ (cf. (23)). The remaining components of the step do not contain deleterious errors; we have

$$\|\widehat{\Delta x}_N\|_\infty = .59(-18), \qquad \|\widehat{\Delta \lambda}\|_\infty = .66(-14), \qquad \|\widehat{\Delta s}_N\|_\infty = .11(-12).$$

Finally, we consider a primal degenerate problem with $|\mathcal{B}| = 4 < m$. The iteration schedule in Table 5 shows similar behavior to the dual degenerate problem. The D factor from iteration 100 is shown in Table 6. All pivots are stable except for the last two 1×1 blocks, which again matches the prediction (63). As discussed in Section 9, the deleterious errors occur in the subvector $\widehat{\Delta \lambda}$, so errors are induced in $\widehat{\Delta s}_N$ and $\widehat{\Delta x}_N$ through formulas 59 and (16b). On iteration 100, we have $\|\widehat{\Delta \lambda}\|_\infty = .32(5)$ and $\|\widehat{\Delta s}_N\|_\infty = .30(7)$ for the affine scaling step. The components $\widehat{\Delta x}_B$ and $\widehat{\Delta s}_B$ are not affected; their ∞-norms are $.17(-18)$ and $.51(-12)$, respectively.

A Proofs of Theorems from Sections 6 and 8

A.1 Proof of Theorem 6.1

We prove (a) by systematically excluding the other possible choices for pivots:

Table 4: The D factor at iteration 17 of the degenerate test problem with $m = 6, n = 12, |\mathcal{B}| = 8$ ($*$ = magnitude less than 10^{-6})

Row/Column	Pivot Block	
1,2	$*$	.95(1)
	.95(1)	$*$
3,4	$*$	-.92(2)
	-.92(2)	$*$
5,6	$*$	.26(2)
	.26(2)	$*$
7	.86(23)	
8	.85(18)	
9	.55(20)	
10	.29(17)	
11,12	$*$	.71(2)
	.71(2)	$*$
13,14	$*$	-.30(0)
	-.30(0)	$*$
15,16	$*$	.15(0)
	.15(0)	$*$
17	.20(-13)	
18	-.60(-19)	

Table 5: Primal degenerate problem: $m = 6, n = 12, |\mathcal{B}| = 4$

k	$\log_{10} \mu_k$	$\log_{10} \|r^k\|_1$	Affine Step?
1	5.4	3.1	
$\vdots$	$\vdots$	$\vdots$	
15	-5.3	-13.9	
16	-8.8	-13.7	$*$
17	-13.7	-14.2	$*$
18	-14.0	-11.6	$*$
$\vdots$	$\vdots$	$\vdots$	
99	-17.6	-13.9	
100	-17.6	-14.0	
$\vdots$	$\vdots$	$\vdots$	

Table 6: The D factor at iteration 17 of the degenerate test problem with $m = 6, n = 12, |\mathcal{B}| = 8$ ($*$ =magnitude less than 10^{-6})

Row/Column	Pivot Block	
1,2	$*$	.95(1)
	.95(1)	$*$
3	.49(23)	
4	.53(19)	
5	.58(19)	
6	.27(20)	
7	.53(9)	
8	.12(21)	
9,10	$*$	.71(2)
	.71(2)	$*$
11,12	$*$	-.15(2)
	-.15(2)	$*$
13	.25(18)	
14	.76(17)	
15,16	$*$	-.16(1)
	-.16(1)	$*$
17	-.15(-8)	
18	-.52(-18)	

(iii) The pivot is 1×1 and is a diagonal element from either the $(1,1)$ or $(2,2)$ blocks of the canonical matrix. Inspection of the Bunch-Kaufman algorithm shows that T_{11} is chosen as pivot if either

$$\chi_1 \leq \frac{|T_{11}|}{\delta} \qquad \text{or} \qquad \chi_1 \leq \sqrt{\frac{\chi_r |T_{11}|}{\delta}}. \tag{64}$$

Now, since χ_r is the maximum off-diagonal in some column of (26), we have $\chi_r = O(1)$, while since T_{11} comes from either the $(1,1)$ or $(2,2)$ block of (26), we have $|T_{11}| = O(\mu + \mathbf{u})$. Since $\delta \in (0,1)$ is fixed, we have from (64) that

$$\chi_1 = O(\mu^{1/2} + \mathbf{u}^{1/2}). \tag{65}$$

Since χ_1 is the magnitude of the largest off-diagonal in some row/column of (26), we have that χ_1 is the ∞-norm of some row or column of B. But (65) is incompatible with $B = O(1)$ and $\kappa(B) = O(1)$. Hence $|T_{11}|$ from the $(1,1)$ or $(2,2)$ blocks cannot be used as a pivot.

A similar argument holds when T_{rr} is chosen as pivot, where T_{rr} is one of the small diagonals.

(iv) The pivot is 2×2 and involves at least one element from Λ. Since all the off-diagonals in (26) are $O(1)$, the quantities χ_i, $i = 1, 2, \ldots, \bar{n}$ are all $O(1)$. A 2×2 pivot with diagonal elements T_{11} and T_{rr} must have

$$|T_{11}| \leq \delta \chi_1, \qquad |T_{rr}| \leq \delta \chi_r,$$

which implies that T_{11} and T_{rr} are both $O(1)$. Since all the diagonals of Λ are $\Omega(\mu^{-1})$, they cannot be candidates for T_{11} and T_{rr}.

(v) The pivot is 2×2, and the pivot block is drawn either entirely from the $(1,1)$ block of (26) or entirely from the $(2,2)$ block. In this case, T_{1r} — the element for which $|T_{1r}| = \chi_1$ — is $O(\mu + \mathbf{u})$. Since T_{1r} has the largest magnitude in its column of (26), and since its column includes either a row or column of B, we have that one of the rows or columns of B is $O(\mu + \mathbf{u})$. As in (iii), we have a contradiction, since this estimate is incompatible with $B = \Omega(1)$ and $\kappa(B) = O(1)$.

This completes the proof of part (a).

We turn to (b), examining the effects of one step of elimination performed with pivot selection corresponding to the two cases (i) and (ii). For (i), suppose the (i,i) element of Λ is chosen as the pivot. After symmetric permutation of the canonical matrix, to place the pivot in the $(1,1)$ position, we obtain

$$\left[\begin{array}{c|c} 1 & \\ \hline & \tilde{P} \end{array}\right] \left[\begin{array}{c|ccc} (\Lambda + O(\mu+\mathbf{u}))_{ii} & N_{\cdot i}^T & 0 & 0 \\ \hline N_{\cdot i} & 0 & B & \tilde{N} \\ 0 & B^T & 0 & 0 \\ 0 & \tilde{N}^T & 0 & \tilde{\Lambda} \end{array}\right] \left[\begin{array}{c|c} 1 & \\ \hline & \tilde{P}^T \end{array}\right] + O(\mu + \mathbf{u}),$$

where $\tilde{P}$ is some permutation matrix, $N_{\cdot i}$ denotes the i-th column of N, $\tilde{N}$ is obtained from N by removing $N_{\cdot i}$, and $\tilde{\Lambda}$ is obtained from Λ by removing its i-th row and column. Since $|(\Lambda + O(\mu + \mathbf{u}))_{ii}^{-1}| = O(\mu)$, the submatrix that remains after elimination is

$$\begin{aligned} \tilde{P} \begin{bmatrix} 0 & B & \tilde{N} \\ B^T & 0 & 0 \\ \tilde{N}^T & 0 & \tilde{\Lambda} \end{bmatrix} \tilde{P}^T - \tilde{P} \begin{bmatrix} N_{\cdot i} \\ 0 \\ 0 \end{bmatrix} \Lambda_{ii}^{-1} \begin{bmatrix} N_{\cdot i}^T & 0 & 0 \end{bmatrix} \tilde{P}^T + O(\mu + \mathbf{u}) \\ = \tilde{P} \begin{bmatrix} 0 & B & \tilde{N} \\ B^T & 0 & 0 \\ \tilde{N}^T & 0 & \tilde{\Lambda} \end{bmatrix} \tilde{P}^T + O(\mu + \mathbf{u}). \end{aligned} \tag{66}$$

It is easy to see that (66) is canonical, so our result is proved for case (i).

For case (ii), the proof is a little messier. Suppose the diagonals of the 2×2 pivot are the (i,i) element of E_1 and the (j,j) element of E_2. After symmetric rearrangement to put this pivot in the upper left corner, (26) becomes

$$\left[\begin{array}{c|c} I & \\ \hline & \hat{P} \end{array}\right] \left[\begin{array}{cc|ccc} 0 & B_{ij} & 0 & B_{i\cdot;j} & N_{i\cdot} \\ B_{ij} & 0 & B_{\cdot j;i}^T & 0 & 0 \\ \hline 0 & B_{\cdot j;i} & 0 & \hat{B} & \hat{N} \\ B_{i\cdot;j}^T & 0 & \hat{B}^T & 0 & 0 \\ N_{i\cdot}^T & 0 & \hat{N}^T & 0 & \Lambda \end{array}\right] \left[\begin{array}{c|c} I & \\ \hline & \hat{P}^T \end{array}\right] + O(\mu + \mathbf{u}),$$

where

- $\hat{P}$ is some permutation matrix;
- $N_{i\cdot}$ is the i-th row of N;
- $\hat{N}$ is N with $N_{i\cdot}$ removed;
- $B_{i\cdot;j}$ is the i-th row of B, with its j-th element removed;
- $B_{\cdot j;i}$ is the j-th column of B, with its i-th element removed;
- $\hat{B}$ is B with its i-th and j-th column removed.

By the choice of B_{ij}, it is either the largest element in its row or the largest element in its column of B. From our assumptions on B, we deduce that $|B_{ij}| = \Omega(1)$. Denoting the pivot block by E, we have

$$E = B_{ij} \begin{bmatrix} 0 & 1 \\ 1 & 0 \end{bmatrix} + O(\mu + \mathbf{u}), \qquad E^{-1} = \frac{1}{B_{ij}} \begin{bmatrix} 0 & 1 \\ 1 & 0 \end{bmatrix} + O(\mu + \mathbf{u}). \tag{67}$$

Therefore the elimination step yields the remaining matrix

$$\begin{aligned} \hat{P} \begin{bmatrix} 0 & \hat{B} & \hat{N} \\ \hat{B}^T & 0 & 0 \\ \hat{N}^T & 0 & \Lambda \end{bmatrix} \hat{P}^T - \tfrac{1}{B_{ij}} \hat{P} \begin{bmatrix} 0 & B_{\cdot j;i} \\ B_{i\cdot;j}^T & 0 \\ N_{i\cdot}^T & 0 \end{bmatrix} \begin{bmatrix} 0 & 1 \\ 1 & 0 \end{bmatrix} \begin{bmatrix} 0 & B_{i\cdot;j} & N_{i\cdot} \\ B_{\cdot j;i}^T & 0 & 0 \end{bmatrix} \hat{P}^T + O(\mu + \mathbf{u}) \\ = \hat{P} \begin{bmatrix} 0 & \hat{B} & \hat{N} \\ \hat{B}^T & 0 & 0 \\ \hat{N}^T & 0 & \Lambda \end{bmatrix} \hat{P}^T + O(\mu + \mathbf{u}), \end{aligned} \tag{68}$$

where

$$\bar{B} = \hat{B} - \frac{1}{B_{ij}} B_{\cdot j;i} B_{i\cdot;j}, \qquad \bar{N} = \hat{N} - \frac{1}{B_{ij}} B_{\cdot j;i} N_{i\cdot}.$$

It is obvious that (68) satisfies Definition 2, except possibly for the conditioning of the remaining matrix $\bar{B}$. This matrix is obtained by pivoting the (i,j) element of B to the $(1,1)$ position and then doing one step of Gaussian elimination. In fact, we are doing partial pivoting since, as noted above, B_{ij} is the largest element in either its row or its column. Hence, the conditioning of the reduced submatrix $\bar{B}$ is unlikely to differ much from $\kappa(B)$, so it is reasonable to assert that $\kappa(\bar{B}) = O(1)$.

We have shown that our stated result holds for both cases (i) and (ii), so our proof of part (b) is complete.

For part (c), note that $C = O(1)$ whether the pivot block is 1×1 or 2×2. For 1×1 pivots, we have $|E| = \Omega(\mu^{-1})$ and $|E^{-1}| = \Omega(\mu)$. For 2×2 pivots, we have from (67) and $|B_{ij}| = \Omega(1)$ that $|E| = O(1)$ and $|E^{-1}| = O(1)$.

A.2 Proof of Theorem 6.2

Again, we prove (a) by excluding the other possible choice for a pivot:

(iii) The pivot is 2×2 and contains at least one element from Λ. In a degenerate canonical matrix, we have $\chi_i = O(\mu + \mathbf{u})$, $i = 1, 2, \ldots, \bar{n}$. A 2×2 pivot with diagonal elements T_{11} and T_{rr} must have

$$|T_{11}| \leq \delta\chi_1, \qquad |T_{rr}| \leq \delta\chi_r,$$

which implies that both diagonals are $O(\mu + \mathbf{u})$, so neither element can come from Λ.

In the case of either a 1×1 or 2×2 pivot made up of elements of size $O(\mu + \mathbf{u})$, we can use the standard argument about element growth in Bunch-Kaufman (that is, the argument that leads to (44)) to deduce the result (b). In the remaining case, where the pivot is a single diagonal element from Λ, we have in the notation of (42) that $|C| = O(\mu + \mathbf{u})$ and $|E| = \Omega(\mu^{-1})$. Hence, the update to the remaining submatrix is bounded by

$$|C||E^{-1}||C|^T = O(\mu(\mu + \mathbf{u})^2),$$

which certainly has size $O(\mu + \mathbf{u})$.

A.3 Proof of Lemma 6.4

Proof. In floating-point arithmetic, the LU factorization of (50) yields the following approximate LU factors:

$$\begin{bmatrix} 1 & 0 \\ E_{11}/E_{12} + \delta_1 & 1 \end{bmatrix}, \qquad \begin{bmatrix} E_{12} & E_{22} \\ 0 & E_{12} - E_{11}E_{22}/E_{12} + \delta_2 \end{bmatrix}, \tag{69}$$

where

$$\delta_1 = \left|\frac{E_{11}}{E_{12}}\right| O(\mathbf{u}), \qquad \delta_2 = |E_{12}|O(\mathbf{u}) + |E_{11}E_{22}/E_{12}|\, O(\mathbf{u}).$$

It is well known that for triangular substitution applied to any triangular system $Uz = h$, the computed solution $\hat{z}$ satisfies $(U + \Delta_U)\hat{z} = h$, where $|E_U| = |U|O(\mathbf{u})$. By applying this observation to each of the matrices in (69), we find that the computed solution $\hat{y}$ of (50) satisfies

$$\begin{bmatrix} 1 & 0 \\ E_{11}/E_{12} + \delta_3 & 1 \end{bmatrix} \begin{bmatrix} E_{12} + \delta_4 & E_{22} + \delta_5 \\ 0 & E_{12} - E_{11}E_{22}/E_{12} + \delta_6 \end{bmatrix} \begin{bmatrix} \hat{y}_1 \\ \hat{y}_2 \end{bmatrix} = \begin{bmatrix} g_2 \\ g_1 \end{bmatrix}, \tag{70}$$

where

$$\begin{aligned}
\delta_3 &= \delta_1 + |E_{11}/E_{12}|\, O(\mathbf{u}) = |E_{11}/E_{12}|\, O(\mathbf{u}), \\
\delta_4 &= |E_{12}|O(\mathbf{u}), \\
\delta_5 &= |E_{22}|O(\mathbf{u}), \\
\delta_6 &= \delta_2 + (|E_{12}| + |E_{11}E_{22}/E_{12}|)\, O(\mathbf{u}) = |E_{12}|O(\mathbf{u}) + |E_{11}E_{22}/E_{12}|\, O(\mathbf{u}).
\end{aligned}$$

By multiplying out the coefficient matrix in (70), we obtain

$$\begin{bmatrix} E_{12}+\delta_4 & E_{22}+\delta_5 \\ E_{11}+\delta_7 & E_{12}+\delta_8 \end{bmatrix}, \tag{71}$$

where

$$\begin{aligned} \delta_7 &= |E_{12}|\delta_3 + |E_{11}/E_{12}|\,\delta_4 = |E_{11}|O(\mathbf{u}), \\ \delta_8 &= |E_{11}/E_{12}|\,\delta_5 + |E_{22}|\delta_3 + \delta_+\left(|E_{12}| + |E_{11}E_{22}/E_{12}|\right)O(\mathbf{u}) \\ &= |E_{11}|O(\mathbf{u}) + |E_{11}E_{22}/E_{12}|\,O(\mathbf{u}) + |E_{12}|O(\mathbf{u}). \\ &= |E_{12}|O(\mathbf{u}). \end{aligned}$$

(The last equality follows from (49).) Hence, (71) can be written as

$$\begin{bmatrix} 0 & 1 \\ 1 & 0 \end{bmatrix}(E+\Delta_E),$$

where Δ_E satisfies the bound (51). ■

A.4 Proof of Theorem 8.1

Proof. We start by proving the analog of Theorem 6.1(a). As in the earlier proof, we systematically exclude the three other possible choices of pivots.

(iii) The pivot is 1×1 and is a diagonal element from either the $(1,1)$ or $(2,2)$ blocks of (26). Then this pivot (T_{ii}, say) will be $O(\mu+\mathbf{u})$. According to the stability criterion (56) we then have $\chi_i = O(\mu+\mathbf{u})$, which implies that one of the rows or columns of B is $O(\mu+\mathbf{u})$. However, this estimate is incompatible with $B = \Omega(1)$ and $\kappa(B) = O(1)$, so this kind of pivot cannot occur.

(iv) The pivot is 2×2 and involves at least one diagonal element from Λ. First, we show that we cannot have both diagonals from Λ. If this were the case, then at least one of these diagonals (T_{ii}, say) would have been considered as a 1×1 pivot at an earlier point in the algorithm. But if it was considered, it would have been accepted, since

$$|T_{ii}^{-1}|\chi_i = O(\mu)O(1) = O(\mu) \le 2/\delta$$

for sufficiently small μ. Hence, at most one of the diagonals is from Λ.

Without loss of generality, suppose in (57) that T_{ii} is from Λ while the remaining diagonal T_{jj} is $O(\mu+\mathbf{u})$. In fact, we have

$$T_{ii} = \Omega(\mu^{-1}), \qquad T_{jj} = O(\mu+\mathbf{u}), \qquad T_{ij} = O(1),$$

and so

$$\left|\begin{bmatrix} T_{ii} & T_{ij} \\ T_{ij} & T_{jj} \end{bmatrix}^{-1}\right| = \frac{1}{|T_{ii}T_{jj} - T_{ij}^2|}\left|\begin{bmatrix} T_{jj} & -T_{ij} \\ -T_{ij} & T_{ii} \end{bmatrix}\right|.$$

Hence, from (57), we have

$$\left|\begin{bmatrix} T_{jj} & -T_{ij} \\ -T_{ij} & T_{ii} \end{bmatrix}\right| \begin{bmatrix} \chi_i \\ \chi_j \end{bmatrix} \leq |T_{ii}T_{jj} - T_{ij}^2| \begin{bmatrix} 1/\delta \\ 1/\delta \end{bmatrix} = \begin{bmatrix} O(1) \\ O(1) \end{bmatrix}.$$

From the second row of this inequality, we have

$$\chi_j \leq \frac{1}{|T_{ii}|} O(1) = O(\mu).$$

But χ_j is the ∞-norm of one of the rows or columns of B, so this estimate contradicts our assumptions on B. Hence, this type of pivot cannot occur.

(v) The pivot is 2×2, and the pivot block E is drawn either entirely from the $(1,1)$ block of (26) or entirely from the $(2,2)$ block. In this case, all elements of E are $O(\mu + \mathbf{u})$. From (57), we have as above that

$$\left|\begin{bmatrix} T_{jj} & -T_{ij} \\ -T_{ij} & T_{ii} \end{bmatrix}\right| \begin{bmatrix} \chi_i \\ \chi_j \end{bmatrix} \leq |T_{ii}T_{jj} - T_{ij}^2| O(1).$$

Taking the second row of this relation, we obtain

$$|T_{ij}|\chi_i + |T_{ii}|\chi_j \leq |T_{ii}T_{jj} - T_{ij}^2| O(1) \leq \left(|T_{ii}T_{jj}| + |T_{ij}|^2\right) O(1), \tag{72}$$

where, by definition, χ_i and χ_j are both nonnegative. Consider two cases. When $|T_{ij}|^2 \geq |T_{ii}T_{jj}|$ we have from (72) that

$$|T_{ij}|\chi_i \leq |T_{ij}|^2 O(1) \implies \chi_i = O(|T_{ij}|) = O(\mu + \mathbf{u}).$$

For the reasons outlined earlier, the assumptions on B are inconsistent with this bound on χ_i, so this case cannot hold. For the other case $|T_{ij}|^2 < |T_{ii}T_{jj}|$, we have

$$|T_{ii}|\chi_j \leq |T_{ii}T_{jj}| O(1) \implies \chi_j = O(|T_{jj}|) = O(\mu + \mathbf{u}),$$

which is also disallowed by our assumptions. Hence, pivots of this type cannot occur.

The proof of the remaining parts (b) and (c) of Theorem 6.1 is identical in this case.

Turning now to the case of a degenerate canonical matrix and the analog of Theorem 6.2, we start by showing that no 2×2 pivots may contain diagonal elements from Λ.

Note that for a degenerate matrix, the off-diagonals, and hence the quantities χ_i, all have size $O(\mu + \mathbf{u})$. If the pivot is a 2×2 block in which both diagonals are from Λ, then one of them (T_{ii}, say) must have been considered previously as a 1×1 pivot. But if it was considered, it would have been accepted, since

$$|T_{ii}^{-1}|\chi_i = O(\mu)O(\mu + \mathbf{u}) \leq 2/\delta.$$

Hence, this type of pivot cannot occur.

If just one of the diagonals is from Λ, this diagonal element (T_{jj}, say) must not have been considered earlier as a 1×1 pivot, since then it would have been accepted for the reason described above. Hence, the other pivot T_{ii}, which has size $O(\mu + \mathbf{u})$, must have been considered as a 1×1 pivot and rejected. Because of (56), T_{ii} must satisfy

$$|T_{ii}| < \frac{\delta}{2}\chi_i. \tag{73}$$

On the other hand, since the 2×2 pivot is accepted, we must have

$$\left|\begin{bmatrix} T_{jj} & -T_{ij} \\ -T_{ij} & T_{ii} \end{bmatrix}\right| \begin{bmatrix} \chi_i \\ \chi_j \end{bmatrix} \leq |T_{ii}T_{jj} - T_{ij}^2| \begin{bmatrix} 1/\delta \\ 1/\delta \end{bmatrix}. \tag{74}$$

Consider first the case of $T_{ij}^2 \geq |T_{ii}T_{jj}|$. Then from the first block row in (74), this inequality implies that

$$|T_{jj}|\chi_i \leq |T_{ii}T_{jj} - T_{ij}^2|\frac{1}{\delta} \leq 2T_{ij}^2\frac{1}{\delta}.$$

Since $|T_{ij}| \leq \chi_i$, we have

$$|T_{jj}| \leq 2|T_{ij}|\frac{1}{\delta} = O(\mu + \mathbf{u}),$$

which contradicts our assumption that T_{jj} has size $\Omega(\mu^{-1})$. The remaining case has $T_{ij}^2 < |T_{ii}T_{jj}|$. From (74) and (73), we have

$$|T_{jj}|\chi_i \leq 2|T_{ii}T_{jj}|\frac{1}{\delta} < 2|T_{jj}|\frac{1}{\delta}\frac{\delta}{2}\chi_i = |T_{jj}|\chi_i,$$

which is a contradiction. Hence this kind of pivot — in which exactly one of the diagonals comes from Λ — cannot occur either, and we are done.

For the analog of part (b) of Theorem 6.2, we have from (56) and (57) and the definition of C and E in (42) that

$$|E^{-1}C^T| \leq |E^{-1}||C^T| = O(1/\delta) = O(1).$$

Hence, the update matrix $CE^{-1}C^T$ is bounded as follows:

$$|CE^{-1}C^T| = \|C\|O(1) = O(\mu + \mathbf{u}),$$

giving the result. ■

References

[1] J. Bunch and L. Kaufman, *Some stable methods for calculating inertia and solving symmetric linear systems*, Mathematics of Computation, 31 (1977), pp. 163–179.

[2] I. S. Duff, *The solution of augmented systems*, Technical Report RAL–93–084, Rutherford Appleton Laboratory, Oxon, U. K., November 1993.

[3] A. Forsgren, P. Gill, and J. Shinnerl, *Stability of symmetric ill-conditioned systems arising in interior methods for constrained optimization*, Report TRITA-MAT-1994-24, Royal Institute of Technology, June 1994.

[4] R. Fourer and S. Mehrotra, *Solving symmetric indefinite systems in an interior-point method for linear programming*, Mathematical Programming, 62 (1993), pp. 15–39.

[5] A. J. Goldman and A. W. Tucker, *Theory of linear programming*, in Linear Equalities and Related Systems, H. W. Kuhn and A. W. Tucker, eds., Primceton University Press, Princeton, N. J., 1956, pp. 53–97.

[6] C. Gonzaga, *Path-following methods in linear programming*, SIAM Review, 34 (1991), pp. 167–224.

[7] O. Güler and Y. Ye, *Convergence behavior of interior-point algorithms*, Mathematical Programming, (1993), pp. 215–228.

[8] M. Kojima, S. Mizuno, and A. Yoshise, *An $O(\sqrt{n}L)$ iteration potential reduction algorithm for linear complementarity problems*, Mathematical Programming, 50 (1991), pp. 331–342.

[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 and Its Applications, 152 (1991), pp. 191–222.

[10] ——, *Computational experience with a globally convergent primal-dual predictor-corrector algorithm for linear programming*, Technical Report SOR 92–10, Program in Statistics and Operations Research, Princeton University, Princeton, N. J., 1992.

[11] ——, *Interior-point methods for linear programming: Computational state of the art*, ORSA Journal on Computing, 6 (1994), pp. 1–14.

[12] N. Megiddo, *On finding primal- and dual-optimal bases*, ORSA Journal on Computing, 3 (1991), pp. 63–65.

[13] S. Mehrotra, *On the implementation of a primal-dual interior point method*, SIAM Journal on Optimization, 2 (1992), pp. 575–601.

[14] S. Mizuno, M. Todd, and Y. Ye, *On adaptive step primal-dual interior-point algorithms for linear programming*, Mathematics of Operations Research, 18 (1993), pp. 964–981.

[15] D. B. Ponceleón, *Barrier methods for large-scale quadratic programming*, PhD thesis, Stanford University, 1990.

[16] M. Todd, *Potential reduction methods in mathematical programming*, Technical Report 1112, Cornell University, Ithaca, N.Y., 14853-3801, 1995.

[17] R. J. Vanderbei, *LOQO User's Manual*, Technical Report SOR 92–5, Program in Statistics and Operations Research, Princeton University, Princeton, N.J., 1992.

[18] S. Vavasis, *Stable numerical algorithms for equilibrium systems*, SIAM Journal on Matrix Analysis and Applications, 15 (1994), pp. 1108–1131.

[19] S. J. Wright, *A path-following infeasible-interior-point algorithm for linear complementarity problems*, Optimization Methods and Software, 2 (1993), pp. 79–106.

[20] ——, *A path-following interior-point algorithm for linear and quadratic optimization problems*, Preprint MCS–P401–1293, Mathematics and Computer Science Division, Argonne National Laboratory, Argonne, Ill., December 1993. (to appear in *Annals of Operations Research*).

[21] ——, *Stability of linear equations solvers in interior-point methods*, Preprint MCS–P400–1293, Mathematics and Computer Science Division, Argonne National Laboratory, Argonne, Ill., December 1993. (to appear in *SIAM Journal on Matrix Analysis and Applications*).

[22] X. Xu, P. Hung, and Y. Ye, *A simplified homogeneous and self-dual linear programming algorithm and its implementation.* Manuscript, September 1993.

[23] Y. Ye, *On the finite convergence of interior-point algorithms for linear programming*, Tech. Rep. 91–5, Department of Management Sciences, University of Iowa, Iowa City, February 1991.

[24] Y. Zhang, *On the convergence of a class of infeasible-interior-point methods for the horizontal linear complementarity problem*, SIAM Journal on Optimization, 4 (1994), pp. 208–227.