STABILITY AND ACCURACY OF INEXACT INTERIOR POINT METHODS FOR CONVEX QUADRATIC PROGRAMMING *

BENEDETTA MORINI§ AND VALERIA SIMONCINI†

Abstract. We consider primal-dual IP methods where the linear system arising at each iteration is formulated in the reduced (augmented) form and solved approximately. Focusing on the iterates close to a solution, we analyze the accuracy of the so-called inexact step, i.e., the step that solves the unreduced system, when combining the effects of both different levels of accuracy in the inexact computation, and different processes for retrieving the step after block elimination. Our analysis is general and includes as special cases sources of inexactness due either to roundoff and computational errors or to the iterative solution of the augmented system using typical procedures. In the roundoff case, we recover and extend some known results.

Key words. primal-dual interior point methods, inexact interior point steps, convex quadratic programming.

1. Introduction. Inexact Interior Point (IP) methods are procedures for linear, convex and nonlinear programming problems where the linear system arising at each iteration is solved approximately. These procedures are mainly associated to the solution of the linear systems by Krylov subspace methods [1, 3–9, 11, 14, 19, 23], but they also provide a framework for analyzing the effect of any source of errors, such as computational errors.

Standard implementations of IP methods use a "reduced" augmented inherently ill-conditioned system obtained by block elimination [13, 27] in spite of a usually well-conditioned system denoted as "unreduced" [15, 20]. This fact is supported by the stability analysis of direct linear equation solvers in IP methods [12, 22, 25, 26, 28, 29], showing that in most cases the computed steps are much more accurate than expected from the general theory on conditioning; see, e.g., [16]. In the context of inexact preconditioned IP methods, the analysis in the recent paper [21] also supports the use of the reduced formulation. In there, a theoretical and computational comparison between unreduced and reduced systems shows that typically the performance of IP implementations with the unreduced and reduced formulations are similar in terms of robustness, while the latter is more convenient in terms of computational efforts due to the smaller dimensions.

This work aims at deepening our understanding of the accuracy in the solution of the unreduced system, i.e., the so-called inexact step, in the late stage of an inexact IP method for convex quadratic programming. The reduced formulation of the linear systems is used, and the inexact step is analyzed considering the effects of both different levels of accuracy in the inexact computation, and different processes proposed in the literature for forming the step after block elimination. A key ingredient in our analysis is the distinction between the basic and nonbasic parts of the step. We provide accuracy results of the inexact step with respect to the "exact" one and establish when near-unit steplengths can be taken eventually by the IP procedure. In case inexactness refers to roundoff-errors, our theory recovers and extends some

*Version of May 4, 2017. Work partially supported by INdAM-GNCS under the 2016 Project *Metodi numerici per problemi di ottimizzazione vincolata di grandi dimensioni e applicazioni.*

†Dipartimento di Matematica, Università di Bologna, Piazza di Porta S. Donato 5, 40127 Bologna, Italia, and IMATI-CNR, Pavia (valeria.simoncini@unibo.it)

§Dipartimento di Ingegneria Industriale, Università degli Studi di Firenze, viale G.B. Morgagni 40, 50134 Firenze, Italia (benedetta.morini@unifi.it).

known results.

In the context of IP methods implemented with iterative Krylov solvers, the application of possibly ill-conditioned preconditioners may also seem to limit the accuracy of the overall solver. We prove that certain commonly employed versions of augmented and constraint preconditioners are *structurally* ill-conditioned, so that if direct solvers are employed within their application, the analysis performed in [12, 25, 26] ensures that the action of the preconditioner does not hinder the stability of the whole procedure.

The paper is organized as follows. In Section 2 we introduce the quadratic optimization problem and we analyze the primal-dual IP method. In Section 3 we explore the inexact solution of the reduced augmented system while in Section 4 we discuss two different ways for recovering the inexact step. In Section 5 we apply the results obtained to the case where the source of inexactness comes from computational errors and relate to the existing literature. In Section 6 we analyze the acceptance of an inexact step and its steplength eventually. In Section 7 we study the spectral properties of two classes of preconditioners for the reduced system and finally in Section 8 we give our conclusions.

Notation. For any vector x, the ith component is denoted as either x_i or $(x)_i$. Given $x \in \mathbb{R}^n$, $X = \mathrm{diag}(x)$ is the diagonal matrix with diagonal entries $x_1, \ldots, x_n$. Given column vectors x and y, we write (x, y) for the column vector given by their concatenation instead of using $[x^T, y^T]^T$. For any positive integer p, the vector of all ones of dimension p and the identity matrix of dimension p are indicated as e_p and I_p, respectively.

For any $x \in \mathbb{R}^n$ and $K \subset \{1, \ldots, n\}$, we write either v_K or $(v)_K$ for the subvector of v having components v_i, $i \in K$. Further, if $A = (a_{ij}) \in \mathbb{R}^{n \times n}$, and $K, L \subset \{1, \ldots, n\}$ we denote either by A_{KL} or $(A)_{KL}$ the submatrix of V with elements a_{ij}, $i \in K$, $j \in L$ and by A_{KK}^{-1} the inverse of A_{KK} when it exists.

The Euclidean vector norm and the associated induced matrix norm are denoted as $\|\cdot\|$.

Given the scalar or vector or matrix v, and the nonnegative scalar χ, we write $v = O(\chi)$ if there is a moderate constant g such that $\|v\| \leq g\chi$. We write $v = \Theta(\chi)$ if there are constants f and g such that $f\chi \leq \|v\| \leq g\chi$.

2. Preliminaries. In this section we introduce the problem considered and the features of the primal-dual IP method adopted for its solution.

We analyze the pair of convex quadratic programming problems formulated as

$$\min\ c^T x + \frac{1}{2} x^T H x \quad \text{subject to} \quad Jx - d = z,\ x \geq 0,\ z \geq 0, \tag{2.1}$$

$$\max\ d^T y - \frac{1}{2} x^T H x \quad \text{subject to} \quad Hx + c - J^T y = s,\ s \geq 0,\ y \geq 0, \tag{2.2}$$

and the pair of convex quadratic programming problems formulated in standard form

$$\min\ c^T x + \frac{1}{2} x^T H x \quad \text{subject to} \quad Jx = d,\ x \geq 0, \tag{2.3}$$

$$\max\ d^T y - \frac{1}{2} x^T H x \quad \text{subject to} \quad J^T y + s - Hx = c,\ s \geq 0, \tag{2.4}$$

where $J \in \mathbb{R}^{m \times n}$ has full row rank $m \leq n$, $H \in \mathbb{R}^{n \times n}$ is symmetric and positive semidefinite, x, s, $c \in \mathbb{R}^n$, y, z, $d \in \mathbb{R}^m$, and the inequalities are meant componentwise. Our analysis is carried out on problems (2.1)–(2.2); the corresponding results for (2.3)–(2.4) are then stated.

The Karush-Kuhn-Tucker (KKT) conditions for (2.1)–(2.2) are of the form

$$\begin{bmatrix} Hx - J^T y - s + c \\ Jx - z - d \\ XSe_n \\ YZe_m \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \\ 0 \\ 0 \end{bmatrix}, \tag{2.5}$$

with $(x, y, s, z) \geq 0$. Primal-Dual IP methods for (2.1)–(2.2) are a modification of the Newton method for (2.5) [27]. They bias the trial steps toward the interior of the nonnegative orthant $(x, y, s, z) \geq 0$, keep all the pairwise products $x_i s_i$ and $y_i z_i$ strictly positive and drive them to zero at the same rate. To this end, the following perturbed KKT conditions are used

$$\begin{bmatrix} Hx - J^T y - s + c \\ Jx - z - d \\ XSe_n \\ YZe_m \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \\ \sigma\mu e_n \\ \sigma\mu e_m \end{bmatrix}, \quad \sigma \in [\sigma_{\min}, \sigma_{\max}] \subset [0, 1], \tag{2.6}$$

where σ is a centering parameter, $X = \text{diag}(x)$, $Y = \text{diag}(y)$, $S = \text{diag}(s)$, $Z = \text{diag}(z)$, and μ is the complementarity gap

$$\mu = \frac{(x, y)^T (s, z)}{n + m}. \tag{2.7}$$

At the kth iteration, IP methods compute the Newton direction and make one step in this direction before reducing the parameter μ. This requires the solution of the linear system

$$\begin{bmatrix} H & -J^T & -I_n & 0 \\ J & 0 & 0 & -I_m \\ S^k & 0 & X^k & 0 \\ 0 & Z^k & 0 & Y^k \end{bmatrix} \begin{bmatrix} \Delta x^k \\ \Delta y^k \\ \Delta s^k \\ \Delta z^k \end{bmatrix} = \begin{bmatrix} -\xi_d^k \\ -\xi_p^k \\ -X^k S^k e_n + \sigma_k \mu_k e_n \\ -Y^k Z^k e_m + \sigma_k \mu_k e_m \end{bmatrix}, \ \sigma_k \in [0, 1]; \tag{2.8}$$

here μ_k is the complementarity gap (2.7) for (x^k, y^k, s^k, z^k), and $\xi^k = (\xi_d^k, \xi_p^k)$ is the residual

$$\xi^k \stackrel{\text{def}}{=} \xi(x^k, y^k, s^k, z^k) = \begin{bmatrix} \xi_d^k \\ \xi_p^k \end{bmatrix} = \begin{bmatrix} Hx^k + c - J^T y^k - s^k \\ Jx^k - z^k - d \end{bmatrix}. \tag{2.9}$$

The standard approach for solving (2.8) is to eliminate the Δs^k and Δz^k variables from the last two equations in (2.8) and solve the reduced augmented system

$$A^k (\Delta_{x,y}^{ex})^k = b^k, \tag{2.10}$$

of the form

$$\begin{bmatrix} H + (X^k)^{-1} S^k & -J^T \\ J & (Y^k)^{-1} Z^k \end{bmatrix} \begin{bmatrix} \Delta x^k \\ \Delta y^k \end{bmatrix} = \begin{bmatrix} -\xi_d^k - s^k + \sigma_k \mu_k (X^k)^{-1} e_n \\ -\xi_p^k - z^k + \sigma_k \mu_k (Y^k)^{-1} e_m \end{bmatrix}. \tag{2.11}$$

A symmetric formulation of this system can be easily obtained using the unknowns $(\Delta x^k, -\Delta y^k)$.

Our analysis in based on assumptions made in the literature. As for the problem data and the solutions we make the same assumptions as in Wright [26]:

ASSUMPTION 2.1.

A1.1 *The vector (x^*, y^*, s^*, z^*) solves (2.1)-(2.2) and strict complementarity holds. The set $\{1, 2, \ldots, n+m\}$ of the indices of the pairs (x^*, y^*) and (s^*, z^*) can be partitioned as*

$$B = \{i \in \{1, 2, \ldots, n+m\} \quad s.t. \quad (x^*, y^*)_i > 0\},$$
$$N = \{1, 2, \ldots, n+m\} \backslash B = \{i \in \{1, 2, \ldots, n+m\} \quad s.t. \quad (s^*, z^*)_i > 0\}.$$

A1.2 *The quantities*

$$\|(x^*, y^*)_B\|, \quad \|\operatorname{diag}((x^*, y^*)_B)^{-1}\|, \quad \|(s^*, z^*)_N\|, \quad \|\operatorname{diag}((s^*, z^*)_N)^{-1}\|,$$

are all of moderate size.

A1.3 *The coefficient matrix in (2.8) is nonsingular at (x^*, y^*, s^*, z^*).*

A1.4 *Letting*

$$M = \begin{bmatrix} H & -J^T \\ J & 0 \end{bmatrix}, \tag{2.12}$$

the quantities $\|M\|$, $\|M_{BB}^{-1}\|$ are of moderate size.

The IP methods considered belong to the class of path-following methods, whose iterates satisfy the following requirements [27].

ASSUMPTION 2.2.

A2.1 *All iterates (x^k, y^k, s^k, z^k) are restricted to neighborhoods of the form*

$$\mathcal{N}_\infty(\gamma) = \left\{(x, y, s, z) > 0 \mid \gamma\,\mu \le (x, y)_i (s, z)_i \le \frac{1}{\gamma}\mu \quad \forall i\right\}, \tag{2.13}$$

for some $\gamma \in (0, 1)$, and μ given in (2.7).

A2.2 *All iterates (x^k, y^k, s^k, z^k) satisfy $\|\xi^k\| \le \beta \mu_k \|\xi^0\|$ for some given positive β.*

The focus of this paper is on the final stage of the IP method and on iterates (x^k, y^k, s^k, z^k) which are close enough to a solution (x^*, y^*, s^*, z^*) satisfying Assumption 2.1. It is known that (x^k, y^k, s^k, z^k) has the following properties.

LEMMA 2.1. *Suppose that Assumptions 2.1–2.2 hold. If the iterate (x^k, y^k, s^k, z^k) is close enough to (x^*, y^*, s^*, z^*), we have*

$$(x^k, y^k)_i \in [\bar{C}_2, \bar{C}_1] \qquad \forall i \in B, \tag{2.14}$$
$$(s^k, z^k)_i \in [\bar{C}_2, \bar{C}_1] \qquad \forall i \in N, \tag{2.15}$$
$$C_1 \mu_k \le (x^k, y^k)_i \le C_2 \mu_k \qquad \forall i \in N, \tag{2.16}$$
$$C_1 \mu_k \le (s^k, z^k)_i \le C_2 \mu_k \qquad \forall i \in B, \tag{2.17}$$

where $\bar{C}_1$, $\bar{C}_2$ are of moderate size, $C_1 = \dfrac{\gamma}{\bar{C}_1}$, $C_2 = \dfrac{1}{\bar{C}_2 \gamma}$.

In addition, b^k given in (2.10)–(2.11) satisfies

$$|(b^k)_i| \le C_3 \mu_k \ \ \forall i \in B, \qquad |(b^k)_i| \le C_3 \ \ \forall i \in N, \tag{2.18}$$

for some C_3, and

$$\|(\Delta x^k, \Delta y^k, \Delta s^k, \Delta z^k)\| = \Theta(\mu_k). \tag{2.19}$$

Proof. See [26, Lemma 2.1] and [26, p. 1292-1294]. □

For our analysis it is convenient to introduce

$$B_x = \{i \in \{1,\dots,n\} \ \text{ s.t. } \ (x^*)_i > 0\}, \qquad N_x = \{1,2,\dots,n\}\backslash B_x, \tag{2.20}$$

$$B_y = \{i \in \{1,\dots,m\} \ \text{ s.t. } \ (y^*)_i > 0\}, \qquad N_y = \{1,2,\dots,m\}\backslash B_y, \tag{2.21}$$

and denote

$$X_B = \mathrm{diag}((x)_{B_x}), \quad Y_B = \mathrm{diag}((y)_{B_y}), \quad X_N = \mathrm{diag}((x)_{N_x}), \quad Y_N = \mathrm{diag}((y)_{N_x}).$$

The above notations are used also for S_B, Z_B, S_N, Z_N, i.e., $S_B = \mathrm{diag}((s)_{B_x})$, $Z_B = \mathrm{diag}((z)_{B_y})$, etc. Consequently, setting $C_4 = \max\left\{\dfrac{\bar{C}_1}{C_1}, \dfrac{C_2}{\bar{C}_2}\right\}$, we get

$$\frac{1}{C_4}\,\mu_k \le \left\| \begin{bmatrix} (X_B^k)^{-1}S_B^k & 0 \\ 0 & (Y_B^k)^{-1}Z_B^k \end{bmatrix} \right\| \le C_4\mu_k, \tag{2.22}$$

$$\frac{1}{C_4}\,\mu_k^{-1} \le \left\| \begin{bmatrix} (X_N^k)^{-1}S_N^k & 0 \\ 0 & (Y_N^k)^{-1}Z_N^k \end{bmatrix} \right\| \le C_4\mu_k^{-1}. \tag{2.23}$$

Remark 1. The use of the symmetric neighborhood (2.2) induced by the infinity norm guarantees that every complementarity product is of order $O(\mu)$. On the other hand, the well-known one-sided infinity neighborhood

$$\mathcal{N}_{-\infty}(\gamma) = \{(x,y,s,z) > 0 \mid (x,y)_i(s,z)_i \ge \gamma\,\mu, \ \text{ for all } \ i = 1,\dots,n+m\}, \tag{2.24}$$

prevents too small complementarity products but some complementarity products may reach a value of order $O((n+m)\mu)$, and this may be a disadvantage in a large-scale setting, see [14].

3. Inexact solution of the augmented system. In practical applications of IP methods, the trial step is computed via the augmented system (2.10). This system becomes ill-conditioned eventually since a subset of diagonal elements tend to become very large in magnitude. Nonetheless, in a series of articles on the accuracy of IP steps it was shown that this type of ill-conditioning can be *benign* [12, 22, 25, 26, 28, 29].

Inspired by the mentioned papers, we consider the case where the augmented system (2.10) is solved approximately, that is there exists a (residual) vector r^k such that

$$A^k(\Delta_{x,y}^{in})^k = b^k + r^k, \tag{3.1}$$

where $(\Delta_{x,y}^{in})^k$ is the computed inexact solution. We are interested in analyzing the effect of the residual vector $r^k \in \mathbb{R}^{n+m}$ on the solution. In this section r^k may represent either the effect of computational errors in IP methods or the effect of an approximate solution of the system [4, 5, 8, 14]. In the subsequent sections we will specialize the origin and the form of r^k; in the cases under study distinct bounds for the subvectors r_B^k and r_N^k will be available.

The inexact step $(\Delta_{x,y}^{in})^k = ((\Delta_x^{in})^k, (\Delta_y^{in})^k)$ and the residual $r^k = (r_x^k, r_y^k)$ in (3.1) satisfy

$$\begin{bmatrix} H + (X^k)^{-1}S^k & -J^T \\ J & (Y^k)^{-1}Z^k \end{bmatrix} \begin{bmatrix} (\Delta_x^{in})^k \\ (\Delta_y^{in})^k \end{bmatrix} = \begin{bmatrix} -\xi_d^k - s^k + \sigma_k\mu_k(X^k)^{-1}e_n \\ -\xi_p^k - z^k + \sigma_k\mu_k(Y^k)^{-1}e_m \end{bmatrix} + \begin{bmatrix} r_x^k \\ r_y^k \end{bmatrix}. \tag{3.2}$$

For ease of notation, from now on we drop the iteration index k in (2.10) and (3.1).

We analyze the systems (2.10) and (3.1) under Assumptions 2.1 and 2.2 and show that the sensitivity of the system to the perturbation r is much smaller than the conditioning of A. To this end, we follow [25, 26] and consider the effect of the perturbation by partitioning A, $\Delta^{in}_{x,y}$, b and r into the two sets B and N introduced in Assumption 2.1.

Applying proper permutations of rows and columns of

$$A = \begin{bmatrix} H + X^{-1}S & -J^T \\ J & Y^{-1}Z \end{bmatrix}, \tag{3.3}$$

and of rows of $\Delta^{ex}_{x,y}$, $\Delta^{in}_{x,y}$, b, r, we write the systems (2.10) and (3.1) as

$$\begin{bmatrix} A_{BB} & A_{BN} \\ A_{NB} & A_{NN} \end{bmatrix} \begin{bmatrix} (\Delta^{ex}_{x,y})_B \\ (\Delta^{ex}_{x,y})_N \end{bmatrix} = \begin{bmatrix} b_B \\ b_N \end{bmatrix}, \tag{3.4}$$

$$\begin{bmatrix} A_{BB} & A_{BN} \\ A_{NB} & A_{NN} \end{bmatrix} \begin{bmatrix} (\Delta^{in}_{x,y})_B \\ (\Delta^{in}_{x,y})_N \end{bmatrix} = \begin{bmatrix} b_B \\ b_N \end{bmatrix} + \begin{bmatrix} r_B \\ r_N \end{bmatrix}, \tag{3.5}$$

where

$$A_{BB} = M_{BB} + \begin{bmatrix} X_B^{-1}S_B & 0 \\ 0 & Y_B^{-1}Z_B \end{bmatrix}, \tag{3.6}$$

$$A_{NN} = M_{NN} + \begin{bmatrix} X_N^{-1}S_N & 0 \\ 0 & Y_N^{-1}Z_N \end{bmatrix}, \tag{3.7}$$

$$A_{BN} = M_{BN},\ A_{NB} = M_{NB}, \tag{3.8}$$

and M is defined in (2.12).

As a first step, we analyze the matrix in (3.4) and its inverse

$$\begin{bmatrix} A_{BB} & A_{BN} \\ A_{NB} & A_{NN} \end{bmatrix}^{-1} = \begin{bmatrix} \Sigma^{-1} & -\Sigma^{-1}A_{BN}A_{NN}^{-1} \\ -A_{NN}^{-1}A_{NB}\Sigma^{-1} & A_{NN}^{-1}(I_{NN} + A_{NB}\Sigma^{-1}A_{BN}A_{NN}^{-1}) \end{bmatrix}, \tag{3.9}$$

with

$$\Sigma = A_{BB} - A_{BN}A_{NN}^{-1}A_{NB}. \tag{3.10}$$

By Assumptions 2.1, 2.2 and by (2.22), (2.23) there exists a scalar C_5 such that

$$\max\{\|A_{BB}\|, \|A_{BN}\|, \|A_{NB}\|\} \le C_5, \quad \|A_{NN}\| \le C_5\mu^{-1}. \tag{3.11}$$

The following assumption is added to our general hypotheses.

ASSUMPTION 3.1. *The iterate (x, y, s, z) is close enough to (x^*, y^*, s^*, z^*) so that Lemma 2.1 holds. The barrier parameter μ is small enough so that*

$$\omega_1 \stackrel{\text{def}}{=} \max\left\{\|M_{NN}\|,\ \|M_{BB}^{-1}\|\left(1 + \frac{C_5^2}{1 - C_4\|M_{NN}\|\mu}\right)\right\} C_4\mu < \frac{1}{2},$$

where C_4 is given in (2.22), and C_5 is given in (3.11).

Our first result shows bounds for the blocks in (3.9). We will repeatedly use the following standard technical lemma.

LEMMA 3.1. *Let $\|\cdot\|$ be any induced matrix norm and $A \in \mathbb{R}^{n\times n}$ be such that $\|A\| < 1$. Then $\|(I+A)^{-1}\| \leq 1/(1-\|A\|)$.*

LEMMA 3.2. *Suppose that Assumptions 2.1, 2.2 and 3.1 hold, and let A given in (3.3) be partitioned as in (3.4) and its inverse as in (3.9). Then*

$$\|A_{NN}^{-1}\| \leq 2C_4\,\mu, \tag{3.12}$$

$$\|\Sigma\| \leq \frac{3}{2}\|M_{BB}\|, \qquad \|\Sigma^{-1}\| \leq 2\|M_{BB}^{-1}\|, \tag{3.13}$$

$$\|\Sigma^{-1}A_{BN}A_{NN}^{-1}\| \leq C_6\mu, . \tag{3.14}$$

with C_4 given in (2.22), C_5 given in (3.11) and C_6 being a positive scalar.

Proof. Note that

$$A_{NN}^{-1} = \begin{bmatrix} X_N^{-1}S_N & 0 \\ 0 & Y_N^{-1}Z_N \end{bmatrix}^{-1} \left(M_{NN} \begin{bmatrix} X_N^{-1}S_N & 0 \\ 0 & Y_N^{-1}Z_N \end{bmatrix}^{-1} + I_{NN} \right)^{-1},$$

and that by (2.23) and Assumption 3.1,

$$\left\| M_{NN} \begin{bmatrix} X_N^{-1}S_N & 0 \\ 0 & Y_N^{-1}Z_N \end{bmatrix}^{-1} \right\| \leq C_4\|M_{NN}\|\mu < \frac{1}{2}.$$

Then, Lemma 3.1 gives

$$\|A_{NN}^{-1}\| \leq 2 \left\| \begin{bmatrix} X_N^{-1}S_N & 0 \\ 0 & Y_N^{-1}Z_N \end{bmatrix}^{-1} \right\|, \tag{3.15}$$

and (3.12) follows from (2.23).

Consider Σ in (3.10),

$$\begin{aligned} \Sigma &= M_{BB} + \begin{bmatrix} X_B^{-1}S_B & 0 \\ 0 & Y_B^{-1}Z_B \end{bmatrix} - A_{BN}A_{NN}^{-1}A_{NB} \\ &= M_{BB}\left(I_{BB} + M_{BB}^{-1}\left(\begin{bmatrix} X_B^{-1}S_B & 0 \\ 0 & Y_B^{-1}Z_B \end{bmatrix} - A_{BN}A_{NN}^{-1}A_{NB} \right) \right) \\ &\stackrel{\text{def}}{=} M_{BB}(I_{BB}+E). \end{aligned} \tag{3.16}$$

Using (2.22) and (3.15) it holds that

$$\|E\| \leq \|M_{BB}^{-1}\| \left(1 + \frac{C_5^2}{1 - C_4\|M_{NN}\|\mu} \right) C_4\mu \leq \omega_1 < \frac{1}{2}.$$

Thus, $\|\Sigma\| \leq \frac{3}{2}\|M_{BB}\|$. From Lemma 3.1, $\|\Sigma^{-1}\| \leq \|M_{BB}^{-1}\|\frac{1}{1-\|E\|}$, and (3.13) follows.

For the off-diagonal blocks in (3.9), clearly (3.14) holds with $C_6 = 4\|M_{BB}^{-1}\|C_4C_5$. ∎

The next proposition shows that the accuracy of $\Delta_{x,y}^{in}$ is higher than expected from inspection of the condition number of A and provides an estimate of the steplength.

PROPOSITION 3.3. *Let Assumptions 2.1, 2.2 and 3.1 hold. Let $\Delta^{ex}_{x,y}$ and $\Delta^{in}_{x,y}$ be as in (3.4) and (3.5). Then there exists a positive scalar C_7 such that*

$$\|(\Delta^{ex}_{x,y} - \Delta^{in}_{x,y})_N\| \le C_7\,\mu(\|r_B\| + \|r_N\|), \tag{3.17}$$
$$\|(\Delta^{in}_{x,y})_N\| \le C_7\mu(1 + \|r_B\| + \|r_N\|), \tag{3.18}$$

and

$$\|(\Delta^{ex}_{x,y} - \Delta^{in}_{x,y})_B\| \le C_7(\|r_B\| + \mu\|r_N\|), \tag{3.19}$$
$$\|(\Delta^{in}_{x,y})_B)\| \le C_7(\mu + \|r_B\| + \mu\|r_N\|). \tag{3.20}$$

Proof. The systems (2.10), (3.1) and equation (3.9) give

$$\begin{aligned}(\Delta^{ex}_{x,y})_N &= -A^{-1}_{NN}A_{NB}\Sigma^{-1}b_B + A^{-1}_{NN}(I_{NN} + A_{NB}\Sigma^{-1}A_{BN}A^{-1}_{NN})b_N,\\ (\Delta^{in}_{x,y})_N &= -A^{-1}_{NN}A_{NB}\Sigma^{-1}(b_B + r_B) +\\ &\quad +A^{-1}_{NN}(I_{NN} + A_{NB}\Sigma^{-1}A_{BN}A^{-1}_{NN})(b_N + r_N),\end{aligned} \tag{3.21}$$

and

$$\begin{aligned}\|(\Delta^{ex}_{x,y})_N - (\Delta^{in}_{x,y})_N\| \le{}& \|A^{-1}_{NN}A_{NB}\Sigma^{-1}\|\,\|r_B\| +\\ &+\|A^{-1}_{NN}\|\,\|I_{NN} + A_{NB}\Sigma^{-1}A_{BN}A^{-1}_{NN}\|\|r_N\|.\end{aligned}$$

Thus, (3.17) follows from (3.11) – (3.14).

In order to derive (3.18) we use (3.21), (3.11) – (3.14) and (2.18). The latter gives $\|b_B\| + \|r_B\| + \|b_N\| + \|r_N\| = O(1 + \|r_B\| + \|r_N\|)$ and implies (3.18) (enlarging C_7 if necessary).

Concerning the set B, the block structure in (3.9) gives

$$(\Delta^{ex}_{x,y})_B = \Sigma^{-1}b_B - \Sigma^{-1}A_{BN}A^{-1}_{NN}b_N, \tag{3.22}$$
$$(\Delta^{in}_{x,y})_B = \Sigma^{-1}(b_B + r_B) - \Sigma^{-1}A_{BN}A^{-1}_{NN}(b_N + r_N). \tag{3.23}$$

Therefore,

$$\|(\Delta^{ex}_{x,y})_B - (\Delta^{in}_{x,y})_B\| \le \|\Sigma^{-1}\|\,\|r_B\| + \|\Sigma^{-1}A_{BN}A^{-1}_{NN}\|\,\|r_N\|,$$

and (3.19) follows from (3.13) and (3.14) (enlarging C_7 if necessary).

Finally, (3.23) and (2.18) give the bound (3.20) (enlarging C_7 if necessary). □

Theorem 3.3 indicates that eventually the part of $\Delta^{in}_{x,y}$ in the set B has an absolute error bound of order $\|r_B\|$ whereas the part of $\Delta^{in}_{x,y}$ in the set N has a much smaller error bound.

4. Recovering the inexact step $(\Delta^{in}_s, \Delta^{in}_z)$. The step $\Delta^{in}_{s,z} = (\Delta^{in}_s, \Delta^{in}_z)$ can be recovered in two ways, either block-wise or element-wise.

I) *Block-wise computation.* The vectors Δ^{in}_s and Δ^{in}_z are formed with a *block-wise* computation based on the first two equations in (2.8). It amounts to setting

$$\Delta^{in}_{s,z} = M\Delta^{in}_{x,y} + \xi, \tag{4.1}$$

with M given in (2.12) and $\xi = (\xi_d, \xi_p)$.

Using (3.2), the inexact step $(\Delta_x^{in}, \Delta_y^{in}, \Delta_s^{in}, \Delta_z^{in})$ satisfies

$$\begin{bmatrix} H & -J^T & -I_n & 0 \\ J & 0 & 0 & -I_m \\ S & 0 & X & 0 \\ 0 & Z & 0 & Y \end{bmatrix} \begin{bmatrix} \Delta_x^{in} \\ \Delta_y^{in} \\ \Delta_s^{in} \\ \Delta_z^{in} \end{bmatrix} = \begin{bmatrix} -\xi_d \\ -\xi_p \\ -XSe_n + \sigma\mu e_n \\ -YZe_m + \sigma\mu e_m \end{bmatrix} + \begin{bmatrix} 0 \\ 0 \\ Xr_x \\ Yr_y \end{bmatrix}. \tag{4.2}$$

Interestingly, if the current iterate is primal and dual feasible, any step along the inexact direction $(\Delta_{x,y}^{in}, \Delta_{s,z}^{in})$ preserves primal and dual feasibility.

II) *Element-wise computation.* The vectors Δ_s^{in} and Δ_z^{in} are formed via the last two equations in (2.8)

$$\Delta_s^{in} = -X^{-1}S\Delta_x^{in} - Se_n + \sigma\mu X^{-1}e_n, \tag{4.3a}$$
$$\Delta_z^{in} = -Y^{-1}Z\Delta_y^{in} - Ze_m + \sigma\mu Y^{-1}e_m. \tag{4.3b}$$

Using again (3.2), the inexact step $(\Delta_x^{in}, \Delta_y^{in}, \Delta_s^{in}, \Delta_z^{in})$ satisfies

$$\begin{bmatrix} H & -J^T & -I_n & 0 \\ J & 0 & 0 & -I_m \\ S & 0 & X & 0 \\ 0 & Z & 0 & Y \end{bmatrix} \begin{bmatrix} \Delta_x^{in} \\ \Delta_y^{in} \\ \Delta_s^{in} \\ \Delta_z^{in} \end{bmatrix} = \begin{bmatrix} -\xi_d \\ -\xi_p \\ -XSe_n + \sigma\mu e_n \\ -YZe_m + \sigma\mu e_m \end{bmatrix} + \begin{bmatrix} r_x \\ r_y \\ 0 \\ 0 \end{bmatrix}. \tag{4.4}$$

Next we compare the step $\Delta_{s,z}^{ex} = (\Delta s, \Delta z)$ in (2.8) with the step $\Delta_{s,z}^{in}$ and distinguish the two cases described above.

PROPOSITION 4.1. *Let Assumptions 2.1, 2.2 and 3.1 hold. Let $\Delta_{x,y}^{ex}$ and $\Delta_{x,y}^{in}$ be as in (3.4) and (3.5) and $\Delta_{s,z}^{in}$ be computed block-wise as in (4.1). Then*

$$\|(\Delta_{s,z}^{ex} - \Delta_{s,z}^{in})_B\| \le C_8(\|r_B\| + \mu\|r_N\|), \tag{4.5}$$
$$\|(\Delta_{s,z}^{ex} - \Delta_{s,z}^{in})_N\| \le C_8(\|r_B\| + \mu\|r_N\|), \tag{4.6}$$

and

$$\|\Delta_{s,z}^{in}\| \le C_8(\mu + \|r_B\| + \mu\|r_N\|), \tag{4.7}$$

where C_8 is a positive scalar.

Proof. By construction $\Delta_{s,z}^{ex} = M\Delta_{x,y}^{ex} + \xi$, and consequently

$$\Delta_{s,z}^{ex} - \Delta_{s,z}^{in} = M(\Delta_{x,y}^{ex} - \Delta_{x,y}^{in}). \tag{4.8}$$

Then bounds (3.17), (3.19) yield to (4.5) and (4.6).

Bound (4.7) follows from (3.18), (3.20) and Assumption 2.2. □

The proof of Proposition 4.1 highlights that the matrix-vector product (4.8) mixes up the vectors $(\Delta_{x,y}^{ex} - \Delta_{x,y}^{in})_B$ and $(\Delta_{x,y}^{ex} - \Delta_{x,y}^{in})_N$ having different accuracy and consequently a component-wise analysis for $\Delta_{s,z}^{in}$ in the sets B and N cannot be carried out.

On the other hand, if $\Delta_{s,z}^{in}$ is computed element-wise via (4.3) then the absolute error bound for $\|(\Delta_{s,z}^{ex} - \Delta_{s,z}^{in})_B\|$ is much smaller than that for $\|(\Delta_{s,z}^{ex} - \Delta_{s,z}^{in})_N\|$. For ease of notation, in the following we use

$$v_{x,y} = (x,\, y), \quad v_{s,z} = (s,\, z).$$

PROPOSITION 4.2. *Let Assumptions 2.1, 2.2 and 3.1 hold. Let $\Delta^{ex}_{x,y}$ and $\Delta^{in}_{x,y}$ be as in (3.4) and (3.5), and $\Delta^{in}_{s,z}$ be computed element-wise as in (4.3). Then*

$$\|(\Delta^{ex}_{s,z} - \Delta^{in}_{s,z})_B\| \le C_8\mu(\|r_B\| + \mu\|r_N\|), \tag{4.9}$$

$$\|(\Delta^{ex}_{s,z} - \Delta^{in}_{s,z})_N\| \le C_8(\|r_B\| + \|r_N\|), \tag{4.10}$$

and

$$\|(\Delta^{in}_{s,z})_B\| \le C_8\mu(1 + \|r_B\| + \|r_N\|), \tag{4.11}$$

$$\|(\Delta^{in}_{s,z})_N\| \le C_8(\mu + \|r_B\| + \|r_N\|), \tag{4.12}$$

$$\|\Delta^{in}_{s,z}\| \le C_8(\mu + \|r_B\| + \|r_N\|), \tag{4.13}$$

where C_8 is a positive scalar.

Proof. By construction

$$\Delta^{ex}_{s,z} = -\begin{bmatrix} X^{-1}S & 0 \\ 0 & Y^{-1}Z \end{bmatrix}\Delta^{ex}_{x,y} - \begin{bmatrix} S & 0 \\ 0 & Z \end{bmatrix}\begin{bmatrix} e_n \\ e_m \end{bmatrix} + \sigma\mu\begin{bmatrix} X^{-1} & 0 \\ 0 & Y^{-1} \end{bmatrix}\begin{bmatrix} e_n \\ e_m \end{bmatrix},$$

$$\Delta^{in}_{s,z} = -\begin{bmatrix} X^{-1}S & 0 \\ 0 & Y^{-1}Z \end{bmatrix}\Delta^{in}_{x,y} - \begin{bmatrix} S & 0 \\ 0 & Z \end{bmatrix}\begin{bmatrix} e_n \\ e_m \end{bmatrix} + \sigma\mu\begin{bmatrix} X^{-1} & 0 \\ 0 & Y^{-1} \end{bmatrix}\begin{bmatrix} e_n \\ e_m \end{bmatrix}.$$

Thus,

$$|(\Delta^{ex}_{s,z} - \Delta^{in}_{s,z})_i| = \left|\frac{(v_{s,z})_i}{(v_{x,y})_i}\right| |(\Delta^{ex}_{x,y} - \Delta^{in}_{x,y})_i|,$$

and the thesis follows from (2.22), (2.23) and Theorem 3.3.

Bounds (4.11)–(4.13) follow from (4.9), (4.10) and (2.19). □

The above results can be used to analyze how the two different strategies for recovering $\Delta^{in}_{s,z}$ affect the steplength behavior of the IP method. We start considering the positivity condition

$$(v_{x,y}, v_{s,z}) + \alpha(\Delta^{in}_{x,y}, \Delta^{in}_{s,z}) > 0, \tag{4.14}$$

and estimate the steplength that can eventually be taken without violating positivity of the iterates. Next theorem refers to the block-wise computation of $\Delta^{in}_{s,z}$. Bound (4.7) for $\|\Delta^{in}_{s,z}\|$ is used and the claim concerns the worst-case analysis for the occurrences where $\alpha = 1$ can be taken eventually. The following two results rely on a large enough magnitude of $\sigma_{\min}$, the left extreme of the interval enclosing the centering parameter σ, as defined in (2.6).

THEOREM 4.3. *Let Assumptions 2.1, 2.2 and 3.1 hold. Let $\Delta^{in}_{x,y}$ be as in (3.5) and $\Delta^{in}_{s,z}$ be computed block-wise via (4.1). If $\sigma_{\min} \in (0,1)$ is sufficiently larger, in a sense to be defined below, than $\gamma\left(\mu + \|r_N\| + \mu^{-1}\|r_B\|\right)$ then condition (4.14) is satisfied for all $\alpha \in [0,1]$.*

Proof. The steplength to the boundary of the positive orthant $(x, y, s, z) > 0$ is affected by the components $\Delta^{in}_{x,y}$ in the set N and the components $\Delta^{in}_{s,z}$ in the set B.

The last two equations in (4.2) give

$$(v_{x,y})_i(\Delta^{in}_{s,z})_i + (v_{s,z})_i(\Delta^{in}_{x,y})_i = -(v_{x,y})_i(v_{s,z})_i + \sigma\mu + (v_{x,y})_i\, r_i. \tag{4.15}$$

Concerning $(\Delta^{in}_{x,y})_i$, $i \in N$, equation (4.15) gives

$$\begin{aligned}(v_{x,y})_i + \alpha(\Delta^{in}_{x,y})_i &= (v_{x,y})_i \left(1 + \alpha \frac{(\Delta^{in}_{x,y})_i}{(v_{x,y})_i}\right) \\ &= (v_{x,y})_i \left(1 - \alpha + \alpha \left(\frac{\sigma\mu}{(v_{x,y})_i (v_{s,z})_i} - \frac{(\Delta^{in}_{s,z})_i}{(v_{s,z})_i} + \frac{r_i}{(v_{s,z})_i}\right)\right) \\ &\geq (v_{x,y})_i \left(1 - \alpha + \alpha \left(\frac{\sigma\mu}{(v_{x,y})_i (v_{s,z})_i} - \left|-\frac{(\Delta^{in}_{s,z})_i}{(v_{s,z})_i} + \frac{r_i}{(v_{s,z})_i}\right|\right)\right).\end{aligned}$$

Using (2.15) and (4.7) we obtain

$$\begin{aligned}\left|-\frac{(\Delta^{in}_{s,z})_i}{(v_{s,z})_i} + \frac{r_i}{(v_{s,z})_i}\right| &\leq \frac{1}{\bar{C}_2}\left(C_8(\|r_B\| + \mu + \mu\|r_N\|) + \|r_N\|\right) \qquad (4.16) \\ &\leq C_9(\mu + \|r_B\| + \|r_N\|), \qquad (4.17)\end{aligned}$$

for some scalar C_9. Using $(x, y, s, z) \in \mathcal{N}_\infty(\gamma)$, we conclude that (4.14) holds for all $\alpha \in [0, 1]$ provided that

$$\sigma\mu > (v_{x,y})_i (v_{s,z})_i \left|-\frac{(\Delta^{in}_{s,z})_i}{(v_{s,z})_i} + \frac{r_i}{(v_{s,z})_i}\right| \geq \gamma\mu \left|-\frac{(\Delta^{in}_{s,z})_i}{(v_{s,z})_i} + \frac{r_i}{(v_{s,z})_i}\right|.$$

Considering the upper bound in (4.17), the previous inequality is guaranteed when $\sigma_{\min} \geq \gamma C_9(\mu + \|r_B\| + \|r_N\|)$.

We next consider $(\Delta^{in}_{s,z})_i$, $i \in B$. Proceeding as above, we have

$$(v_{s,z})_i + \alpha(\Delta^{in}_{s,z})_i \geq (v_{s,z})_i \left(1 - \alpha + \alpha \left(\frac{\sigma\mu}{(v_{x,y})_i (v_{s,z})_i} - \left|-\frac{(\Delta^{in}_{x,y})_i}{(v_{x,y})_i} + \frac{r_i}{(v_{s,z})_i}\right|\right)\right),$$

and (2.14), (2.17) and (3.20) give

$$\begin{aligned}\left|-\frac{(\Delta^{in}_{x,y})_i}{(v_{x,y})_i} + \frac{r_i}{(v_{s,z})_i}\right| &\leq \frac{C_7}{\bar{C}_2}(\mu + \|r_B\| + \mu\|r_N\|) + \frac{1}{C_1\mu}\|r_B\| \\ &\leq C_9(\mu + \mu^{-1}\|r_B\| + \mu\|r_N\|),\end{aligned}$$

(enlarging C_9 if necessary). This concludes the proof. □

Now, consider the case where $\Delta^{in}_{s,z}$ is computed element-wise.

THEOREM 4.4. *Let Assumptions 2.1, 2.2 and 3.1 hold. Let $\Delta^{in}_{x,y}$ be as in (3.5) and $\Delta^{in}_{s,z}$ be computed via (4.3). If $\sigma_{\min} \in (0, 1)$ is sufficiently larger, in a sense to be defined below, than $\gamma(\mu + \|r_B\| + \|r_N\|)$, then condition (4.14) is satisfied for all $\alpha \in [0, 1]$.*

Proof. The steplength to the boundary of the positive orthant $(x, y, s, z) > 0$ is affected by the components $\Delta^{in}_{x,y}$ in the set N and the components $\Delta^{in}_{s,z}$ in the set B. The proof parallels that in Lemma 4.3. By (4.3),

$$(v_{x,y})_i (\Delta^{in}_{s,z})_i + (v_{s,z})_i (\Delta^{in}_{x,y})_i = -(v_{x,y})_i (v_{s,z})_i + \sigma\mu. \qquad (4.18)$$

For the components $\Delta^{in}_{x,y}$ in the set N, it holds

$$(v_{x,y})_i + \alpha(\Delta^{in}_{x,y})_i \geq (v_{x,y})_i \left(1 - \alpha + \alpha \left(\frac{\sigma\mu}{(v_{x,y})_i(v_{s,z})_i} - \left|\frac{(\Delta^{in}_{s,z})_i}{(v_{s,z})_i}\right|\right)\right).$$

By using (2.15) and (4.13), we conclude that condition (4.14) is guaranteed for all $\alpha \in [0,1]$, provided that $\sigma_{\min}$ is sufficiently larger than $\gamma(\mu + \|r_B\| + \|r_N\|)$.

As for $(\Delta^{in}_{s,z})_i$, $i \in B$, we have

$$(v_{s,z})_i + \alpha(\Delta^{in}_{s,z})_i \geq (v_{s,z})_i \left(1 - \alpha + \alpha \left(\frac{\sigma\mu}{(v_{x,y})_i(v_{s,z})_i} - \left|-\frac{(\Delta^{in}_{x,y})_i}{(v_{x,y})_i}\right|\right)\right).$$

Then (2.14) and (3.20) imply that (4.14) is guaranteed for all $\alpha \in [0,1]$, provided that $\sigma_{\min}$ is sufficiently larger than $\gamma(\mu + \|r_B\| + \mu\|r_N\|)$.

Combining the results for the set B and N, the proof is completed. □

Comparing the claims in Theorem 4.3 and Theorem 4.4 we observe that, when $\Delta^{in}_{s,z}$ is computed via the block-wise strategy in (4.1), the term $\mu^{-1}\|r_B\|$ may increase as μ tends to zero and the inexact step may be severely damped in a long-step implementation of the IP method. While the result for the element-wise computation is in the same flavor as that in [28], the result for the block-wise strategy appears to be new.

5. Computational errors in IP methods. One possible source of inexactness in the step computation derives from floating point arithmetic with precision $\mathbf{u}$. The analysis given in the previous section can be used to characterize the behavior of IP methods in the presence of computational errors and in this section we address such an issue. The result stated for the element-wise computation of the step recovers that in [26, 28], while the result concerning the block-wise computation of the step is new.

Let $\Delta^{in}_{x,y}$ be an approximation to the step $\Delta^{ex}_{x,y}$ with perturbation caused by roundoff errors in the solution of (2.11). Let $\sigma \in [\sigma_{\min}, \sigma_{\max}] \subset [0,1]$, $\rho \in (0,1)$ and consider a standard IP scheme where the new iterate is formed finding the largest $\alpha \in [0,1]$ such that

$$(v_{x,y}, v_{s,z}) + \alpha(\Delta^{in}_{x,y}, \Delta^{in}_{s,z}) \in \mathcal{N}_\infty(\gamma), \tag{5.1}$$

$$\mu(\alpha) \stackrel{\text{def}}{=} \frac{(v_{x,y} + \alpha\Delta^{in}_{x,y})^T(v_{s,z} + \alpha\Delta^{in}_{s,z})}{n+m} \geq (1-\alpha)\mu, \tag{5.2}$$

$$\mu(\alpha) \leq \rho\mu. \tag{5.3}$$

see [27, p. 111]. In [26] S. Wright studied how roundoff errors affect the IP method and showed when the computed steps are sufficiently accurate to be useful. Specifically, suppose that $\Delta^{in}_{x,y}$ is computed by applying Gaussian elimination to (2.11) and suppose that $\Delta^{in}_{s,z}$ is computed element-wise , i.e. by (4.3). Then roundoff errors do not badly affect the IP method as unit steplengths can be taken eventually. Indeed, Theorem 4.3 in [26] claims that the steplength parameter α is one if σ_{min} is substantially larger that $(\mu + \mathbf{u})$. This result relies on the following issues shown in [26, Theorems 3.2, 3.3, 3.4]: the computed step can be viewed as an inexact step $\Delta^{in}_{x,y}$ which satisfies (3.2) with $(r_B, r_N) = O(\mu + \mathbf{u})$; it holds

$$\begin{aligned} &\|(\Delta^{ex}_{x,y} - \Delta^{in}_{x,y})_N\| = O(\mu(\mu+\mathbf{u})), \quad \|(\Delta^{ex}_{s,z} - \Delta^{in}_{s,z})_B\| = O(\mu(\mu+\mathbf{u})), \\ &\|(\Delta^{ex}_{x,y} - \Delta^{in}_{x,y})_B\| = O(\mu+\mathbf{u}), \quad \|(\Delta^{ex}_{s,z} - \Delta^{in}_{s,z})_N\| = O(\mu+\mathbf{u}), \end{aligned}$$

and

$$\|(\Delta^{in}_{x,y})_N\| = O(\mu), \quad \|(\Delta^{in}_{x,y})_B\| = O(\mu + \mathbf{u}), \tag{5.4}$$

$$\|(\Delta^{in}_{s,z})_B\| = O(\mu), \quad \|(\Delta^{in}_{s,z})_N\| = O(\mu + \mathbf{u}). \tag{5.5}$$

Remarkably, for the components of the iterate which tend to become active, the steplength to the boundary is not affected by finite precision arithmetic.

We note that the above estimates can be derived also from our Theorems 3.3 and 4.2 setting $(r_B, r_N) = O(\mu + \mathbf{u})$. As for problems (2.3)-(2.4), in accordance to [28, Theorems 4.1, 5.3], we obtain

$$\|(\Delta x)_N\| = O(\mu), \quad \|(\Delta x)_B\| = O(\mu + \mathbf{u}), \tag{5.6}$$

$$\|(\Delta s)_B\| = O(\mu), \quad \|(\Delta s)_N\| = O(\mu + \mathbf{u}). \tag{5.7}$$

A similar result does not hold when $\Delta^{in}_{s,z}$ is computed block-wise. Neglecting for simplicity computational errors in the computation of $\Delta^{in}_{s,z}$, by Theorem 4.1 we have

$$\|\Delta^{in}_{s,z}\| = O(\mu + \mathbf{u}), \tag{5.8}$$

and distinct estimates in the sets B and N, such as those in (5.5), are not available. Consequently, for μ small enough $\|\Delta^{in}_{s,z}\| = O(\mathbf{u})$ and roundoff errors may be amplified as μ decreases. Indeed, considering the basic requirement of positivity of the iterates, Theorem 4.3 gives that, in the most adverse case, the steplength is affected by the quantity $\mu^{-1}\|r_B\| = O(\mu^{-1}(\mu + \mathbf{u}))$ which is essentially of order $O(\mu^{-1}\mathbf{u})$ if μ is smaller than $\mathbf{u}$. As for problems (2.3)-(2.4) in standard form, proceeding as above bound (5.6) and (5.8) hold.

5.1. Numerical illustration of the step behavior. We report on typical behavior of the IP step, supporting our findings of Section 5. To focus on finite precision arithmetic, the linear systems are solved by Gaussian elimination with partial pivoting. The constants in the IP method are set as: $\sigma = \text{mid}(10^{-2}, \mu/\sqrt{n}, 2.\,10^{-1})$, $\gamma = 10^{-3}$, the attainable maximum steplength α is reduced by the factor 0.995. In order to analyze asymptotic effects, as in [28] termination occurs with the artificially stringent control $\mu \leq 10^{-30}$.

Our experiments indicate that the formula used for recovering $\Delta^{in}_{s,z}$ does not influence the IP iterations as long as μ and ξ are reduced to $O(\mathbf{u})$. Successively, for stringent values of μ we know that $\|\Delta^{in}_{s,z}\| = O(\mathbf{u})$ when the step is computed block-wise, and $\|(\Delta^{in}_{s,z})_B\| = O(\mu)$ when the step is computed element-wise, and that in the former case the steplength may be severely shortened. This occurrence has been detected in some runs and here we report a representative case concerning problem LPnetlib/lp_israel [24]. The problem is in standard form, its dimensions are $n = 316$, $m = 174$, and $J \in \mathbb{R}^{m\times n}$ is full row rank. Primal and dual regularization is applied and the quadratic regularization terms are of the form $10^{-6}I_n$ and $10^{-6}I_m$.

In Table 5.1 we report selected iterations in the late stage of the IP method and display: the value of μ; the maximum value of α in $(0, 1]$ satisfying (5.1)–(5.3) in case $\Delta^{in}_{s,z}$ is computed element-wise; the maximum value of α in $(0, 1]$ required for positivity of the iterates in case $\Delta^{in}_{s,z}$ is computed element-wise. Moreover, we display the components s_i, $(\Delta s)_i$ of s and Δs corresponding to the minimum value $(\Delta s)_i/s_i$ attained for $i = 1, \ldots, n$; observe that if $\min_i (\Delta s)_i/(s_i) > -1$, positivity is guaranteed with $\alpha = 1$. The notation $w(q)$ means $w \cdot 10^q$. Remarkably, for very small values of μ, the element-wise computation of $\Delta^{in}_{s,z}$ allows steps with maximal (unit) steplength, whereas block-wise computation leads to significantly smaller steps.

	$\Delta^{in}_{s,z}$ computed element-wise			$\Delta^{in}_{s,z}$ computed block-wise		
μ	α	s_i	$(\Delta s)_i$	α	s_i	$(\Delta s)_i$
1.6(−17)	1.0	1.4642(−17)	−1.4478(−17)	1.0	1.4642(−17)	−1.4478(−17)
3.5(−21)	1.0	3.6102(−21)	−3.5736(−21)	1.0	6.3711(−22)	−6.3067(−22)
1.2(−26)	1.0	1.2200(−26)	−1.2078(−26)	9.7(−1)	3.7759(−27)	−3.8654(−27)
2.6(−30)	1.0	2.7277(−30)	−2.7004(−30)	6.6(−3)	8.4425(−31)	−1.2622(−28)

TABLE 5.1

Selected IP iterations: complementarity gap μ, maximum value of α allowed using element-wise computation for $\Delta^{in}_{s,z}$; maximum value of α allowed for positivity using block-wise computation for $\Delta^{in}_{s,z}$.

6. Inexact Interior Point methods. In this section we consider IP methods where the large linear system arising at each iteration is solved by an iterative method, giving rise to an inexact procedure. The key issue in this context is the level of error acceptable in the approximation of the steps, see, e.g., [5, 8, 14].

Two ways for controlling the residual $r = (r_x, r_y)$, $r_x \in \mathbb{R}^n$, $r_y \in \mathbb{R}^m$, in (3.2) have been devised. They are based on the theory in [10] for Inexact Newton methods and on the observation that the right-hand side in (2.8) is of order $O(\mu)$. If $\Delta^{in}_{s,z}$ is computed block-wise then the step $(\Delta^{in}_x, \Delta^{in}_y, \Delta^{in}_s, \Delta^{in}_z)$ solves (4.2) and the residual is required to satisfy the following stopping criterion

$$\left\| \begin{bmatrix} X & 0 \\ 0 & Y \end{bmatrix} r \right\| = \left\| \begin{bmatrix} Xr_x \\ Yr_y \end{bmatrix} \right\| \leq \eta\mu, \quad \eta \in (0,1); \tag{6.1}$$

see, e.g., [8, 14]. Since a scaling is applied on the residual r, we name (6.1) the Scaled Residual (SRE) criterion.

If $\Delta^{in}_{s,z}$ is formed via element-wise computation, then $(\Delta^{in}_x, \Delta^{in}_y, \Delta^{in}_s, \Delta^{in}_z)$ solves (4.4) and the residual is requested to satisfy

$$\|r\| \leq \eta\mu, \quad \eta \in (0,1); \tag{6.2}$$

see, e.g., [4–6]. We denote this control as NonScaled Residual (NOSRE) criterion.

The accuracy of the inexact steps is as follows. We recall that the centering parameter σ belongs to the interval $[\sigma_{\min}, \sigma_{\max}] \subset (0,1)$.

COROLLARY 6.1. *Let Assumptions 2.1, 2.2 and 3.1 hold. Let $(\Delta^{in}_{x,y}, \Delta^{in}_{s,z})$ be the inexact step in (4.2) and r satisfy the* SRE *criterion (6.1). Then there exists a positive scalar C_{10} such that*

$$\|(\Delta^{ex}_{x,y} - \Delta^{in}_{x,y})_N\| \leq C_{10}\,\mu\eta, \qquad \|(\Delta^{ex}_{x,y} - \Delta^{in}_{x,y})_B\| \leq C_{10}\,\mu\eta,$$
$$\|(\Delta^{ex}_{s,z} - \Delta^{in}_{s,z})_B\| \leq C_{10}\,\mu\eta, \qquad \|(\Delta^{ex}_{s,z} - \Delta^{in}_{s,z})_N\| \leq C_{10}\,\mu\eta.$$

If $\sigma_{\min}$ is sufficiently larger than $\gamma(\mu+\eta)$ then condition (4.14) is satisfied for all $\alpha \in [0,1]$.

Proof. Bound (2.14) gives

$$\bar{C}_2\|r_B\| \leq \min_i\{(x_i)_B, (y_i)_B\}\|r_B\| \leq \left\| \begin{bmatrix} Xr_x \\ Yr_y \end{bmatrix}_B \right\| \leq \eta\mu,$$

i.e, $\|r_B\| \leq \dfrac{1}{\bar{C}_2}\,\eta\mu$. Analogously, bound (2.16) gives $\|r_N\| \leq \dfrac{1}{C_1}\,\eta$. Then, the claims are a consequence of Theorems 3.3, 4.1 and 4.3. □

COROLLARY 6.2. *Let Assumptions 2.1, 2.2 and 3.1 hold. Let* $(\Delta^{in}_{x,y}, \Delta^{in}_{s,z})$ *be the inexact step in (4.4) and* r *satisfy the* NOSRE *criterion (6.2). Then there exists a positive scalar* C_{10} *such that*

$$\|(\Delta^{ex}_{x,y} - \Delta^{in}_{x,y})_N\| \le C_{10}\,\mu^2\eta \quad \|(\Delta^{ex}_{x,y} - \Delta^{in}_{x,y})_B\| \le C_{10}\,\mu\eta,$$
$$\|(\Delta^{ex}_{s,z} - \Delta^{in}_{s,z})_B\| \le C_{10}\,\mu^2\eta, \quad \|(\Delta^{ex}_{s,z} - \Delta^{in}_{s,z})_N\| \le C_{10}\,\mu\eta.$$

If $\sigma_{\min}$ *is sufficiently larger than* $\gamma\mu$*, with* γ *as in Assumption 2.2, then condition (4.14) is satisfied for all* $\alpha \in [0,1]$.

Proof. The criterion (6.2) implies $\|r_B\| \le \eta\mu$ and $\|r_N\| \le \eta\mu$, and Theorems 3.3, 4.2 and 4.4 give the claims. □

We next analyze the late stage of a inexact IP method step. We first consider the case where the SRE criterion is used. The new iterate $(v_{x,y}, v_{s,z}) + \alpha(\Delta^{in}_{x,y}, \Delta^{in}_{s,z})$, is formed with the largest $\alpha \in (0,1]$ such that conditions (5.1)–(5.3) are met. As for condition (5.2), it makes the primal and dual infeasibility ξ bounded by some multiple of μ, as required in Assumption 2.2; in fact, the definition of ξ in (2.9) and the first two equations in (4.2) give

$$\xi((v_{x,y}, v_{s,z}) + \alpha(\Delta^{in}_{x,y}, \Delta^{in}_{s,z})) = (1-\alpha)\xi(v_{x,y}, v_{s,z}), \tag{6.3}$$

and this, along with (5.2), implies the requirement on ξ, see [27, p. 111].

THEOREM 6.3. *Let Assumptions 2.1, 2.2 and 3.1 hold. Let* $(\Delta^{in}_{x,y}, \Delta^{in}_{s,z})$ *be the inexact step in (4.2) and* r *satisfy the* SRE *criterion (6.1). If* $\sigma_{\min} \in (0,1)$ *is sufficiently larger, in a sense to be defined below, than* $\dfrac{\eta+\mu}{1-\gamma}$ *and* $\sigma_{\max} \in (0,1)$ *is such that* $\sigma_{\max} < \dfrac{2\rho}{3-\gamma}$*, then conditions (5.1)–(5.3) are satisfied with* α *equal to one.*

Proof. Equation (4.15) gives

$$\begin{aligned}(v_{x,y} + \alpha\Delta^{in}_{x,y})_i(v_{s,z} + \alpha\Delta^{in}_{s,z})_i &= (v_{x,y})_i(v_{s,z})_i + \alpha\left((v_{x,y})_i(\Delta^{in}_{s,z})_i + \right.\\ &\qquad \left.(v_{s,z})_i(\Delta^{in}_{x,y})_i\right) + \alpha^2(\Delta^{in}_{x,y})_i(\Delta^{in}_{s,z})_i\\ &= (1-\alpha)(v_{x,y})_i(v_{s,z})_i + \alpha(\sigma\mu + (v_{x,y})_i r_i + \alpha(\Delta^{in}_{x,y})_i(\Delta^{in}_{s,z})_i).\end{aligned}$$

By summing over i and recalling the definition of $\mu(\alpha)$ in (5.2) we get

$$\mu(\alpha) = \mu(1-\alpha+\alpha\sigma) + \frac{\alpha}{n+m}(v^T_{x,y}r + \alpha(\Delta^{in}_{x,y})^T\Delta^{in}_{s,z}).$$

Observe that (6.1) gives $|(v_{x,y})_i r_i| \le \eta\mu$, $\forall i$, i.e., $\dfrac{|v^T_{x,y}r|}{n+m} \le \eta\mu$. Further, letting $\theta_1 = \|\Delta^{in}_{x,y}\|\,\|\Delta^{in}_{s,z}\|$, it trivially follows $|(\Delta^{in}_{x,y})_i(\Delta^{in}_{s,z})_i| \le \theta_1$ and $\left|\dfrac{(\Delta^{in}_{x,y})^T\Delta^{in}_{s,z}}{n+m}\right| \le \theta_1$. Then, using $(x,y,s,z) \in \mathcal{N}_\infty(\gamma)$, we obtain

$$\begin{aligned}(v_{x,y} + \alpha\Delta^{in}_{x,y})_i(v_{s,z} + \alpha\Delta^{in}_{s,z})_i &\le (1-\alpha)\frac{1}{\gamma}\mu + \alpha(\sigma\mu + |(v_{x,y})_i r_i + \alpha(\Delta^{in}_{x,y})_i(\Delta^{in}_{s,z})_i|)\\ &\le (1-\alpha)\frac{1}{\gamma}\mu + \alpha\left(\sigma\mu + \eta\mu + \theta_1\right), \qquad (6.4)\\ (v_{x,y} + \alpha\Delta^{in}_{x,y})_i(v_{s,z} + \alpha\Delta^{in}_{s,z})_i &\ge (1-\alpha)\gamma\mu + \alpha(\sigma\mu - |(v_{x,y})_i r_i + \alpha(\Delta^{in}_{x,y})_i(\Delta^{in}_{s,z})_i|)\\ &\ge (1-\alpha)\gamma\mu + \alpha\left(\sigma\mu - \eta\mu - \theta_1\right), \qquad (6.5)\end{aligned}$$

and

$$\mu(\alpha) \leq \mu(1-\alpha+\alpha\sigma) + \frac{\alpha}{n+m}|v_{x,y}^T r + \alpha(\Delta_{x,y}^{in})^T \Delta_{s,z}^{in}|$$
$$\leq \mu\left(1-\alpha+\alpha\left(\sigma+\eta+\frac{\theta_1}{\mu}\right)\right), \tag{6.6}$$
$$\mu(\alpha) \geq \mu(1-\alpha+\alpha\sigma) - \frac{\alpha}{n+m}|v_{x,y}^T r + \alpha(\Delta_{x,y}^{in})^T \Delta_{s,z}^{in}|$$
$$\geq \mu\left(1-\alpha+\alpha\left(\sigma-\eta-\frac{\theta_1}{\mu}\right)\right). \tag{6.7}$$

Then, (6.5) and (6.6) give

$$\frac{(v_{x,y}+\alpha\Delta_{x,y}^{in})_i (v_{s,z}+\alpha\Delta_{s,z}^{in})_i}{\mu(\alpha)} \geq \frac{(1-\alpha)\gamma + \alpha(\sigma-\eta-\frac{\theta_1}{\mu})}{1-\alpha+\alpha(\sigma+\eta+\frac{\theta_1}{\mu})},$$

whereas (6.4) and (6.7) give

$$\frac{(v_{x,y}+\alpha\Delta_{x,y}^{in})_i (v_{s,z}+\alpha\Delta_{s,z}^{in})_i}{\mu(\alpha)} \leq \frac{(1-\alpha)\frac{1}{\gamma} + \alpha(\sigma+\eta+\frac{\theta_1}{\mu})}{1-\alpha+\alpha(\sigma-\eta-\frac{\theta_1}{\mu})}.$$

As a consequence, (5.1) is guaranteed for all $\alpha \in [0,1]$ when

$$\sigma_{\min} \geq \frac{2}{1-\gamma}\left(\eta+\frac{\theta_1}{\mu}\right). \tag{6.8}$$

Using (6.7) we can conclude that the above choice of $\sigma_{\min}$ enforces (5.2) too.

Moreover, (6.8) and $\sigma_{\max} < \dfrac{2\rho}{3-\gamma}$ yield

$$\frac{\mu(1)}{\mu} \leq \sigma+\eta+\frac{\theta_1}{\mu} \leq \sigma+\frac{\sigma_{\min}}{2}(1-\gamma) \leq \frac{\sigma(3-\gamma)}{2} \leq \rho.$$

Observing that Corollary 6.1 gives $\theta_1 = O\left(\mu^2\right)$, the proof is completed. □

REMARK 2. *It can be easily verified that the claim in the above theorem is valid also when the* `SRe` *control on the residual is performed using the infinity norm (as occurs in [14]).*

We now estimate the steplength that can eventually be taken when the `NoSRe` criterion is used. Due to the first two equations in (4.4), equality (6.3) does not hold and we control the infeasibility by means of

$$\|\xi((v_{x,y}, v_{s,z}) + \alpha(\Delta_{x,y}^{in}, \Delta_{s,z}^{in}))\| \leq \tau\mu(\alpha), \tag{6.9}$$

$\tau = \beta\|(\xi_d^0, \xi_p^0)\|/\mu^0$, $\beta \geq 1$, see e.g., [27, p. 109]. Summarizing, the new iterate is supposed to meet conditions (5.1), (5.3) and (6.9).

THEOREM 6.4. *Let Assumptions 2.1, 2.2 and 3.1 hold. Let $(\Delta_{x,y}^{in}, \Delta_{s,z}^{in})$ be the inexact step in (4.4) and r satisfy (6.2). If $\sigma_{\min} \in (0,1)$ is sufficiently larger, in a sense to be defined below, than $\dfrac{\mu}{1-\gamma} + \dfrac{\eta}{\tau}$ and $\sigma_{\max} \in (0,1)$ is such that $\sigma_{\max} < \dfrac{2\rho}{3-\gamma}$, then conditions (5.1), (5.3) and (6.9) are satisfied with α equal to one.*

Proof. The proof follows the lines of the proof of Theorem 6.3. Equation (4.18) gives

$$(v_{x,y} + \alpha\Delta^{in}_{x,y})_i(v_{s,z} + \alpha\Delta^{in}_{s,z})_i = (1-\alpha)(v_{x,y})_i(v_{s,z})_i + \alpha(\sigma\mu + \alpha(\Delta^{in}_{x,y})_i(\Delta^{in}_{s,z})_i).$$

and (6.4)–(6.7) become

$$\begin{aligned}
(v_{x,y} + \alpha\Delta^{in}_{x,y})_i(v_{s,z} + \alpha\Delta^{in}_{s,z})_i &\leq (1-\alpha)\frac{1}{\gamma}\mu + \alpha\left(\sigma\mu + \theta_1\right), \\
(v_{x,y} + \alpha\Delta^{in}_{x,y})_i(v_{s,z} + \alpha\Delta^{in}_{s,z})_i, &\geq (1-\alpha)\gamma\mu + \alpha\left(\sigma\mu - \theta_1\right), \\
\mu(\alpha) &\leq \mu\left(1 - \alpha + \alpha\left(\sigma + \frac{\theta_1}{\mu}\right)\right), \\
\mu(\alpha) &\geq \mu\left(1 - \alpha + \alpha\left(\sigma - \frac{\theta_1}{\mu}\right)\right),
\end{aligned}$$

where $\theta_1 = \|\Delta^{in}_{x,y}\|\,\|\Delta^{in}_{s,z}\|$. Then, (5.1) holds if $\sigma_{\min} \geq \dfrac{2}{1-\gamma}\,\dfrac{\theta_1}{\mu}$, and (5.3) is fulfilled if $\sigma_{\max} < \dfrac{2\rho}{3-\gamma}$.

Finally, (4.4) and (6.2) give

$$\tau\mu(1) - \|\xi((v_{x,y}, v_{s,z}) + (\Delta^{in}_{x,y}, \Delta^{in}_{s,z}))\| = \tau\mu(1) - \|(r_x, r_y)\| \geq \tau\mu\left(\sigma - \frac{\theta_1}{\mu} - \frac{\eta}{\tau}\right),$$

so that (5.1) and (6.9) are satisfied if $\sigma_{\min}$ is larger than $\dfrac{2}{1-\gamma}\,\dfrac{\theta_1}{\mu} + \dfrac{\eta}{\tau}$. Since $\theta_1 = O(\mu^2)$ by Corollary 6.2, the proof is completed. □

REMARK 3. *If the* `NoSRe` *control on the residual is performed using the infinity norm, the claim in the above theorem is still valid if σ_{min} is sufficiently larger than* $\dfrac{\mu}{1-\gamma} + \sqrt{n+m}\,\dfrac{\eta}{\tau}$.

We conclude this section drawing some conclusions from Sections 5 and 6. Excluding computational errors, Theorems 6.3 and 6.4 indicate that, for μ small enough so that $\mu \leq \eta$, unit steps can be taken if $\sigma_{\min}$ is of order $O(\eta)$. On the other hand, if the inexact steps are computed in finite precision, the results in Theorems 6.3 and 6.4 are valid as long as roundoff errors in Krylov solvers are below the threshold $\eta\mu$ in both the SRE and NOSRE criteria; otherwise, we fall in the case $r = O(\mathbf{u})$ and the analysis in Section 5 holds.

7. Preconditioners for the augmented system. The system solution phase is also critical for the method overall accuracy. Stability of direct methods for linear systems arising in IP methods was investigated in [12, 22, 25, 26, 28, 29]. Such papers cover the solution of both the augmented linear system (2.11) and the "condensed" system, i.e. the reduced positive definite normal equation formulation of (2.11). Although such systems are increasingly ill-conditioned as the solution is approached, it has been repeatedly reported that this type of ill-conditioning is *benign* when specific direct methods are applied. Recommendations of such analysis are: backward stable methods are used; the corresponding growth factor is bounded by a reasonable constant; the IP iterates are well-centered within the positivity orthant; see, e.g., [12, 25, 27].

The indefinite matrix in (2.10) is diagonally ill-conditioned, i.e., a subset of the diagonal elements becomes very large in magnitude [12]. As reviewed in Section 5, S. Wright [26] shows that the numerical solution of the augmented linear system (2.11) by Gaussian elimination with partial pivoting gives sufficiently accurate computed steps; an *unsymmetric* formulation of the system is used in [26]. Analogous conclusions were drawn by Forsgren et al., [12] for a symmetric and *indefinite* formulation. In there, an LDL^T symmetric factorization is used, where a pivoting strategy selects the large pivots first and, once the dominant rows and columns have been eliminated, the factorization continues on the remaining matrices with either 1×1 and 2×2 pivots as necessary. The key feature for obtaining the above results is that nonbasic indices are eventually used as pivots before any of the basic indices are used, because of the magnitude of the diagonal entries of $X_N^{-1}S_N$, $Y_N^{-1}Z_N$. For these formulations, stability of the Bunch-Parlett, Bunch-Kaufman and sparse Bunch-Parlett algorithms for (2.10) was studied in [28] making use of the fact that generally the largest diagonal elements are not selected as pivots before any others. Useful steps are obtained with nondegenerate linear programming problems.

For the *symmetric positive definite* condensed formulation, the linear system matrices are structurally ill-conditioned, i.e., the eigenvalues fall into two well-separated groups and the submatrices associated to such partitioning are much better conditioned than the whole matrix itself. M. Wright [25] shows that in this case ill-conditioning only marginally affects the accuracy of the computed step.

In the context of *Inexact* IP methods, the approximate system solution is obtained by an iterative method, usually in the Krylov subspace class, accelerated with a preconditioning strategy. In the following we assume that the roundoff of the Krylov iteration itself is under control (see, e.g., [18] for a thorough account), and focus on the preconditioning step. Since the preconditioner should somehow mimic the coefficient matrix, we expect it to also be ill-conditioned. Thus the question arises whether we can rely on some *structured* ill-conditioning for the selected preconditioner, so that if direct solves are involved in its application, known theory ensures good stability properties. In the following we answer in the affirmative for two commonly employed classes of preconditioners: augmented block diagonal and constraint preconditioners. The literature cited above can thus be exploited to justify a good practical behavior of the preconditioners for μ close to zero in spite of their ill-conditioning.

We start by discussing the use of the augmented preconditioner

$$P_A = \begin{bmatrix} P_1 & 0 \\ 0 & P_2 \end{bmatrix} = \begin{bmatrix} H + J^T W^{-1} J + X^{-1} S & 0 \\ 0 & W + Y^{-1} Z \end{bmatrix}, \tag{7.1}$$

where P_1 and P_2 are symmetric and positive definite, W is diagonal positive definite, see e.g., [20]. Typically, W is either a multiple of the identity or it is related to the parameter matrices. We note that $J^T W^{-1} J$ is low rank, and this affects the lower bounds involving this matrix. Assume first that $W = \delta I_m$ with $\delta \gg \mu$. Then for $\|\zeta\| = 1$ we have

$$0 \leq \zeta^T J^T W^{-1} J \zeta \leq \frac{1}{\delta} \|J\|^2. \tag{7.2}$$

Consider next the ideal choice $W = Y^{-1}Z$ [20] and recall the definitions (2.20), (2.21). Applying proper permutations of rows and columns, split W as the block diagonal matrix $W = \text{blkdiag}(W_{B_y}, W_{N_y})$ and the rows of J accordingly, i.e., $J = \begin{bmatrix} J_{B_y:} \\ J_{N_y:} \end{bmatrix}$, with $J_{B_y:}$ and $J_{N_y:}$ being the submatrices of J with row indices in B_y and

N_y respectively. We have $J^T W^{-1} J = J_{B_y:}^T W_{B_y}^{-1} J_{B_y:} + J_{N_y:}^T W_{N_y}^{-1} J_{N_y:}$, so that for $\|\zeta\| = 1$,

$$0 \leq \zeta^T J^T W^{-1} J \zeta \leq C_4 \frac{1}{\mu} \|J_{B_y:}\|^2 + C_4 \mu \|J_{N_y:}\|^2. \tag{7.3}$$

To derive helpful bounds for the spectral properties of P_1 we recall the following result in our context.

THEOREM 7.1. *[17, Theorem 4.2.10] Let $A \in \mathbb{R}^{n \times n}$ be symmetric, with its eigenvalues arranged in increasing order, let $\mathcal{S}$ be a given k-dimensional subspace of $\mathbb{R}^n$, and let $c \in \mathbb{R}$ be given. Then*

i) If $x^T A x \geq c$ for every unit vector $x \in \mathcal{S}$, then $\lambda_{n-k+1}(A) \geq c$.

ii) If $x^T A x \leq c$ for every unit vector $x \in \mathcal{S}$, then $\lambda_k(A) \leq c$.

We can thus derive the following result. We denote with $\mathcal{B}_x$ and $\mathcal{N}_x$ the spaces associated with the indices in B_x and N_x, respectively. To make the result as simple as possible, the theorem analyzes the generic case, that is we assume that $\mathcal{B}_x$, $\mathcal{N}_x$ and $\mathcal{B}_y$, $\mathcal{N}_y$ are non-empty. A few additional comments covering some of the non-generic cases are reported in Remark 4.

PROPOSITION 7.2. *Suppose that (x, y, s, z) satisfies (2.14) – (2.17). Assume that $\mathcal{B}_x$, $\mathcal{N}_x$ and $\mathcal{B}_y$, $\mathcal{N}_y$ are non-empty. With the previous notation,*

i) If $W = \delta I_n$ with $\delta \gg \mu$ then there are $card(\mathcal{B}_x)$ eigenvalues λ of P_1 satisfying $\lambda \leq C^* \left(1 + \frac{1}{\delta} + \mu\right)$ for some positive constant C^*, and $card(N_x)$ eigenvalues that satisfy $\frac{C_*}{\mu} \leq \lambda$ for some positive constant C_*.

ii) Let $\mathcal{S} = null(J) \cap \mathcal{B}_x$ and let n_S be its dimension. If $W = Y^{-1} Z$ then there are n_S eigenvalues λ of P_1 that satisfy $\lambda \leq C^*(1 + \mu)$ for some positive constant C^*, and $n - n_S$ ones that satisfy $\frac{C_*}{\mu} \leq \lambda$ for some positive constant C_*.

Proof. *i)* We write $P_1 = H + \frac{1}{\delta} J^T J + X^{-1} S$. For $\zeta \in \mathcal{B}_x$, $\|\zeta\| = 1$ and using (2.22), (7.2) we obtain

$$\zeta^T P_1 \zeta \leq \lambda_{\max}(H) + \frac{1}{\delta} \|J\|^2 + C_4 \mu.$$

The first result thus follows from Theorem 7.1(*ii*). For $\zeta \in \mathcal{N}_x$, $\|\zeta\| = 1$ and using (2.23) we obtain

$$\lambda_{\min}(H) + \frac{1}{C_4} \frac{1}{\mu} \leq \zeta^T P_1 \zeta.$$

The second result thus follows from Theorem 7.1(*i*).

ii) Let $\zeta \in \mathcal{S}$, $\|\zeta\| = 1$. Then $\zeta^T P_1 \zeta = \zeta^T H \zeta + \zeta^T X^{-1} S \zeta \leq \zeta^T H \zeta + C_4 \mu$. The first result follows from Theorem 7.1. Let now $\zeta \in \mathcal{S}^\perp$, the space orthogonal to $\mathcal{S}$, and $\|\zeta\| = 1$. Then if $\zeta \in \text{null}(J)^\perp$, because of the contribution from $J^T W^{-1} J$ and of (2.23), we obtain $\zeta^T P_1 \zeta \geq \zeta^T J^T W^{-1} J \zeta \geq \frac{1}{C_4} \frac{1}{\mu} \|J\zeta\|^2$ (here we are using the assumption that $\mathcal{B}_y$ is non-empty).

Analogously, if $\zeta \in \text{null}(J)$ and $\zeta \notin \mathcal{B}_x$, because of the contribution from $X^{-1}S$ we obtain $\zeta^T P_1 \zeta \geq \zeta^T X^{-1} S \zeta \geq \dfrac{1}{C_4}\dfrac{1}{\mu}\zeta^T\zeta$ (here we are using the assumption that $\mathcal{B}_x$ is non-empty). The result follows from Theorem 7.1. □

REMARK 4. If $\mathcal{B}_x = \emptyset$ then $\mathcal{S} = \emptyset$ and Proposition 7.2(ii) ensures that all eigenvalues behave like $O\left(\dfrac{1}{\mu}\right)$. The case $\mathcal{B}_y = \emptyset$ is a little more involved and it requires digging into the proof of Proposition 7.2. If $\mathcal{B}_y$ is the empty set, then $\|J^T W^{-1} J\| \leq \dfrac{1}{C_4}\mu\|J\|^2$. Then if $\zeta \in \text{null}(J)^\perp$ and $\zeta \in \mathcal{B}_x$, then $\zeta^T P_1 \zeta = \zeta^T H \zeta + \zeta^T J^T W^{-1} J \zeta + \zeta^T X^{-1} S \zeta \leq \zeta^T H \zeta + C\mu$. If $\zeta \in \text{null}(J)^\perp$ and $\zeta \notin \mathcal{B}_x$, then $\zeta^T P_1 \zeta \geq \zeta^T X^{-1} S \zeta \geq \dfrac{1}{C_4}\dfrac{1}{\mu}$. Summarizing, with respect to Proposition 7.2(ii) we have $card(B_x)$ additional eigenvalues λ of P_1 satisfying $\lambda \leq C^*(1+\mu)$ while the remaining eigenvalues are $O\left(\dfrac{1}{\mu}\right)$.

The analysis of the (2,2) block, $P_2 = W + Y^{-1}Z$ is easier because for both choices of W, P_2 is diagonal. For $W = \delta I$ the block is structurally ill-conditioned in the sense that there is a basic part that is controlled by $\dfrac{1}{\delta} + \mu$, and a nonbasic part that behaves like $\dfrac{1}{\mu}$. Therefore, there will be two eigenvalue clusters, one depending on the magnitude of δ and one on that of μ. Note however that solving with P_2 causes no round-off propagation, since P_2 is diagonal. Hence this ill-conditioning may be considered harmless.

Proposition 7.2 shows that, for both choices of W, matrix P_1 is structurally ill-conditioned similarly to the condensed matrices arising in the late stage of an IP method. Indeed, the first block P_1 has two widely separated sets of "moderate" and "large" eigenvalues. Following [25], if the right-hand side of a system with matrix P_1 is perturbed then the effective conditioning of the problem is split into the conditioning of the two clusters. More generally, using the theory in [25] (see equations (6.9), (6.10) therein) we can deduce that if a backward stable method is used to solve with P_1 then the portion of the computed step in the nonbasic indices has a much smaller error bound than the basic part.

The second class of preconditioners we consider is written as

$$P_C := \begin{bmatrix} D & J^T \\ J & -Y^{-1}Z \end{bmatrix}, \quad D = \text{diag}(H + X^{-1}S);$$

it is an indefinite matrix and is usually called a *constraint* preconditioner in certain contexts where the blocks enforce the constraints stemming from the problem. As outlined at the beginning of this section, the application of P_C is stable using Gaussian elimination with pivoting [26] or the LDL^T factorizations with 1×1 and 2×2 diagonal blocks analyzed in [12, 28]. Alternatively P_C can be factorized as

$$P_C = \begin{bmatrix} I & 0 \\ JD^{-1} & I \end{bmatrix} \begin{bmatrix} D & 0 \\ 0 & -JD^{-1}J^T - Y^{-1}Z \end{bmatrix} \begin{bmatrix} I & D^{-1}J^T \\ 0 & I \end{bmatrix},$$

or, to avoid the occurrence of JD^{-1}, in the equivalent form

$$P_C = \mathcal{L}\mathcal{D}\mathcal{L}^T \equiv \begin{bmatrix} D & 0 \\ J & I \end{bmatrix} \begin{bmatrix} D^{-1} & 0 \\ 0 & -JD^{-1}J^T - Y^{-1}Z \end{bmatrix} \begin{bmatrix} D & J^T \\ 0 & I \end{bmatrix}.$$

Applying P_C requires two block triangular solves with matrices having diagonal blocks, and one block diagonal solve. From [16] we know that the accuracy in the solution of a block triangular system depends on both the conditioning of the diagonal matrices and the norm of the non-diagonal block. Since the non-diagonal block is bounded, the latter decomposition of P_C is preferable. Solving with the (1,1) diagonal block of $\mathcal{D}$ may be considered harmless while solving with the (2,2) block of $\mathcal{D}$ requires further study. Before we derive actual estimates, we emphasize that in spite of the indefiniteness of P_C, the system we need to worry about is symmetric and definite. If we manage to show structured spectral clustering of the eigenvalues of the (2,2) block, then the discussion performed for P_A carries over, and the available theory is again applicable.

The following result ensures that under certain hypotheses the spectrum of the (2,2) block of $\mathcal{D}$ is bounded by moderate constants from below and from above.

PROPOSITION 7.3. *Suppose that (x, y, s, z) satisfies (2.14) – (2.17). With the previous notation, the following holds*

i) There are $card(B_y)$ eigenvalues λ of the (2,2) block of $\mathcal{D}$ satisfying

$$C_*\mu \leq \lambda \leq C^* \left((\min_i h_{ii})^{-1} + \mu \right) \tag{7.4}$$

for some positive constants C_* and C^*;

ii) There are $card(N_y)$ eigenvalues satisfying $C_*\mu^{-1} \leq \lambda$;

iii) If $J^T_{:B_x}$ is full-column rank then the lower bound in (7.4) can be replaced by $C_* \leq \lambda$.

Proof. The matrix $JD^{-1}J^T$ is full rank for any positive μ. Applying proper permutations of rows and columns, split D as the block diagonal matrix $D = \text{blkdiag}(D_{B_x}, D_{N_x})$ and the rows of J accordingly; then $JD^{-1}J^T = J_{:B_x}D^{-1}_{B_x}J^T_{:B_x} + J_{:N_x}D^{-1}_{N_x}J^T_{:N_x}$. Using (2.22) and (2.23), for $\|\zeta\| = 1$ and for some positive scalar C we have

$$\zeta^T JD^{-1}J^T\zeta \leq C \left(\min_i (h_{ii} + \mu)^{-1} \|J^T_{:B_x}\|^2 + \min_i (h_{ii} + \mu^{-1})^{-1} \|J^T_{:N_x}\|^2 \right),$$

together with

$$\begin{aligned}
\zeta^T JD^{-1}J^T\zeta &\geq C \left(\max_i (h_{ii} + \mu)^{-1} \|J^T_{:B_x}\zeta\|^2 + \max_i (h_{ii} + \mu^{-1})^{-1} \|J^T_{:N_x}\zeta\|^2 \right) \\
&\geq C \min \left\{ \max_i (h_{ii} + \mu)^{-1}, \max_i (h_{ii} + \mu^{-1})^{-1} \right\} \|J^T\zeta\|^2 \\
&\geq C \min \left\{ \max_i (h_{ii} + \mu)^{-1}, \max_i (h_{ii} + \mu^{-1})^{-1} \right\} \sigma_{\min}(J),
\end{aligned}$$

where $H = (h_{i,j})$ and $\sigma_{\min}(J)$ is the smallest singular value of J. Thus, for μ sufficiently small the eigenvalues of $JD^{-1}J^T$ satisfy

$$\widetilde{C}\mu \leq \lambda(JD^{-1}J^T) \leq \widehat{C} \left(\frac{1}{\min_i h_{ii}} + \mu \right),$$

for some positive $\widetilde{C}$ and $\widehat{C}$. Moreover, if $J^T_{:B_x}$ is full-column rank the lower bound becomes

$$\widetilde{C} \frac{1}{\max_i (h_{ii} + \mu)} \leq \lambda(JD^{-1}J^T).$$

Using again the bounds in (2.22) and (2.23), for the general case we can deduce that

$$\left(\widetilde{C}\mu + \frac{1}{C_4}\mu\right) \leq \zeta^T(JD^{-1}J^T + Y^{-1}Z)\zeta \leq \left(\widehat{C}\left(\frac{1}{\min_i h_{ii}} + \mu\right) + C_4\mu\right), \quad \zeta \in \mathcal{B}_y$$
$$\left(\widetilde{C}\mu + \frac{1}{C_4}\frac{1}{\mu}\right) \leq \zeta^T(JD^{-1}J^T + Y^{-1}Z)\zeta, \quad \zeta \in \mathcal{N}_y,$$

and, in particular, if $J^T_{:B_x}$ is full-column rank

$$\left(\widetilde{C}\frac{1}{\max_i(h_{ii}+\mu)} + \frac{1}{C_4}\mu\right) \leq \zeta^T(JD^{-1}J^T + Y^{-1}Z)\zeta \leq \left(\widehat{C}\left(\frac{1}{\min_i h_{ii}} + \mu\right) + C_4\mu\right), \quad \zeta \in \mathcal{B}_y$$
$$\left(\widetilde{C}\frac{1}{\max_i(h_{ii}+\mu)} + \frac{1}{C_4}\frac{1}{\mu}\right) \leq \zeta^T(JD^{-1}J^T + Y^{-1}Z)\zeta, \quad \zeta \in \mathcal{N}_y.$$

□

REMARK 5. The spectrum of $JD^{-1}J^T + Y^{-1}Z$ is characterized by a structured clustering depending on μ and $\min_i h_{ii}$. If $\min_i h_{ii}$ is nonzero, which is the case if regularization is employed [2], then there are $card(N_y)$ eigenvalues becoming large as μ tends to zero. Otherwise the large eigenvalues in magnitude are expected to be more than $card(N_y)$.

REMARK 6. In case $J^T_{:B_x}$ is full-column rank and $\max_i h_{ii}$ is nonzero the small eigenvalues are bounded away from zero. The full-column rank hypothesis on $J^T_{:B_x}$ is typically related to the Linear Independence Constraint Qualification (LICQ). We refer to [15, 20] for connection between LICQ and the properties of the unreduced system formulation in IP methods for convex quadratic programming.

8. Conclusions. We have focused on the late phase of an inexact primal-dual interior point method and analyzed when the inexact step is accurate enough to preserve the features of the "exact" step. The sources of inexactness have encompassed both computational errors and the approximation associated with the system solves arising at each iteration. We have shown that unit inexact steps can be taken - hence improving the duality measure - in spite of a possibly severe ill-conditioning of the associated linear system. Moreover, we have shown that the preconditioning phase also provides a less ill-conditioned solution procedure than expected by the global spectral properties of the involved matrices.

REFERENCES

[1] G. Al-Jeiroudi and J. Gondzio. Convergence analysis of inexact infeasible interior point method for linear optimization. *Optimization Methods and Software*, 141:231–247, 2009.

[2] A. Altman and J. Gondzio. Regularized symmetric indefinite systems in interior point methods for linear and quadratic optimization. *Optimization Methods and Software*, 11:275–302, 1999.

[3] P. Armand, J. Benoist, and J.P. Dussault. Local path-following property of inexact interior methods in nonlinear programming. *Computational Optimization and Applications*, 52:209–238, 2012.

[4] V. Baryamureeba and T. Steihaug. On the convergence of an inexact primal-dual interior point method for linear programming. In I. Lirkov, S. Margenov, and J. Wasniewski, editors, *Proceedings of the 5th International Conference on Large-Scale Scientific Computing*, Lecture Notes in Computer Science 3743, pages 629–637, Berlin, 2006. Springer-Verlag.

[5] S. Bellavia. Inexact interior point method. *Journal of Optimization Theory and Applications*, 96:109–121, 1998.

[6] L. Bergamaschi, J. Gondzio, and G. Zilli. Preconditioning indefinite systems in interior point methods for optimization. *Computational Optimization and Applications*, 28:149–171, 2004.

[7] S. Cafieri, M. D'Apuzzo, V. De Simone, and D. di Serafino. On the iterative solution of KKT systems in potential reduction software for large-scale quadratic problems. *Computational Optimization and Applications*, 38:27–45, 2007.

[8] S. Cafieri, M. D'Apuzzo, V. De Simone, and D. di Serafino. Stopping criteria for inner iterations in inexact potential reduction methods: a computational study. *Computational Optimization and Applications*, 36:165–193, 2007.

[9] M. D'Apuzzo, V. De Simone, and D. di Serafino. On mutual impact of linear algebra and large-scale optimization with focus on interior point methods. *Computational Optimization and Applications*, 45:283–310, 2010.

[10] R.S. Dembo, S.C. Eisenstat, and T. Steihaug. Inexact Newton methods. *SIAM Journal on Numerical Analysis*, 19:400–408, 1982.

[11] C. Durazzi and V. Ruggiero. Global convergence of the Newton interior-point method for nonlinear programming. *Journal of Computational Optimization and Applications*, 120:199–208, 2004.

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

[13] A. Forsgren, P.E. Gill, and M.H. Wright. Interior methods for nonlinear optimization. *SIAM Review*, 443:525–597, 2002.

[14] J. Gondzio. Convergence analysis of an inexact feasible interior point method for convex quadratic programming. *SIAM Journal of Optimization*, 23:1510–1527, 2013.

[15] C. Greif, E. Moulding, and D. Orban. Bounds on eigenvalues of matrices arising from interior-point methods. *SIAM Journal on Optimization*, 24:49–83, 2014.

[16] N.J. Higham. *Accuracy and stability of numerical algorithms*. SIAM, Philadelphia, USA, 1996.

[17] R. A. Horn and C. R. Johnson. *Matrix Analysis*. Cambridge University Press, Cambridge, second edition, 2013.

[18] Joerg Liesen and Zdenek Strakos. *Krylov Subspace Methods. Principles and Analysis*. Oxford University Press, 2012.

[19] S. Mizuno and F. Jarre. Global and polynomial-time convergence of an infeasible-interior-point algorithm using inexact computations. *Mathematical Programming*, 84:105–122, 1999.

[20] B. Morini, V. Simoncini, and M. Tani. Spectral estimates for unreduced symmetric KKT systems arising from interior point methods. *Numerical Linear Algebra with Applications*, 23:776–800, 2016.

[21] B. Morini, V. Simoncini, and M. Tani. A comparison of reduced and unreduced KKT systems arising from interior point methods. *Computational Optimization and Applications*, published online:DOI: 10.1007/s10589–017–9907–8, 2017.

[22] D.B. Poncelon. Barrier methods for large-scale quadratic programming. *Technical report*, SOL 91-2, 1991.

[23] T. Rees. The iterative solution of linear systems arising in the primal-dual interior point algorithm for linear programming. *RAL Preprint RAL-P-2016-007*, 2016.

[24] The University of Florida Sparse Matrix Collection. http://www.cise.ufl.edu/research/sparse/matrices/.

[25] M.H. Wright. Ill-conditioning and computational error in interior methods for nonlinear programming. *SIAM Journal on Optimization*, 9:84–111, 1998.

[26] S.J. Wright. Stability of linear equations solvers in interior-point methods. *SIAM Journal on Matrix Analysis and Applications*, 16:1287–1307, 1995.

[27] S.J. Wright. *Primal-dual interior-point methods*. SIAM, Philadelphia, USA, 1997.

[28] S.J. Wright. Stability of augmented system factorizations in interior-point methods. *SIAM Journal on Matrix Analysis and Applications*, 18:191–222, 1997.

[29] S.J. Wright. Effects of finite-precision arithmetic on interior-point methods for nonlinear programming. *SIAM Journal on Optimization*, 12:36–78, 2001.