

# On the convergence of iterative schemes for solving a piecewise linear system of equations

Nicolas F. Armijo <sup>\*</sup>      Yunier Bello-Cruz<sup>†</sup>      Gabriel Haeser <sup>‡</sup>

January 19, 2023

## Abstract

This paper is devoted to studying the global and finite convergence of the semi-smooth Newton method for solving a piecewise linear system that arises in cone-constrained quadratic programming problems and absolute value equations. We first provide a negative answer via a counterexample to a conjecture on the global and finite convergence of the Newton iteration for symmetric and positive definite matrices. Additionally, we discuss some surprising features of the semi-smooth Newton iteration in low dimensions and its behavior in higher dimensions. Secondly, we present two iterative schemes inspired by the classical Jacobi and Gauss-Seidel methods for linear systems of equations for finding a solution to the problem. We study sufficient conditions for the convergence of both proposed procedures, which are also sufficient for the existence and uniqueness of solutions to the problem. Lastly, we perform some computational experiments designed to illustrate the behavior (in terms of CPU time) of the proposed iterations versus the semi-smooth Newton method for dense and sparse large-scale problems. Moreover, we included the numerical solution of a discretization of the Boussinesq PDE modeling a two-dimensional flow in a homogeneous phreatic aquifer.

**Keywords:** Piecewise linear system, quadratic programming, semi-smooth Newton method.

**2010 AMS Subject Classification:** 90C33, 15A48.

## 1 Introduction

We consider the following piecewise linear system:

$$x^+ + Tx = b, \tag{1}$$

where, denoting by  $\mathbb{R}^{n \times n}$  the set of  $n \times n$  matrices with real entries, the data consists of a vector  $b \in \mathbb{R}^n \equiv \mathbb{R}^{n \times 1}$  and a nonsingular matrix  $T \in \mathbb{R}^{n \times n}$ . The variable  $x = (x_1, x_2, \dots, x_n)^T$  is a vector in  $\mathbb{R}^n$  and  $x^+$  is the projection of  $x$  onto  $\mathbb{R}_+^n$ , which has the  $i$ -th component equal to  $x_i^+ = \max\{x_i, 0\}$ ,  $i = 1, \dots, n$ . Some works dealing with problem (1) and its generalizations include

---

<sup>\*</sup>Department of Applied Mathematics, University of São Paulo, Brazil (e-mail: [nfarmijo@ime.usp.br](mailto:nfarmijo@ime.usp.br)). The author was supported by Fapesp grant 2019/13096-2.

<sup>†</sup>Northern Illinois University, USA (e-mail: [yunierbello@niu.edu](mailto:yunierbello@niu.edu)). The author was supported in part by R&A Grant from NIU.

<sup>‡</sup>Department of Applied Mathematics, University of São Paulo, Brazil (e-mail: [ghaeser@ime.usp.br](mailto:ghaeser@ime.usp.br)). The author was supported by CNPq and Fapesp grant 2018/24293-0.

[1–7]. Solutions of equation (1) are closely related to at least two important classes of well-known problems, such as the quadratic cone-constrained programming:

$$\begin{aligned} & \text{minimize} && \frac{1}{2}x^T Qx + q^T x, \\ & \text{subject to} && x \in \mathbb{R}_+^n, \end{aligned} \tag{2}$$

and the absolute value equation [5, 8]:

$$\hat{T}x - |x| = \hat{b}. \tag{3}$$

Namely, the projection onto  $\mathbb{R}_+^n$  of a solution of problem (1) with  $T = (Q - \text{Id})^{-1}$  and  $b = Tq$  satisfies the linear complementarity problem given by the first order optimality conditions of (2) (see [9] for details), while, on the other hand, for  $\hat{T} = -2T - \text{Id}$ , and  $\hat{b} = -2b$ , noting that  $x^+ = \frac{x+|x|}{2}$  one can see that problems (1) and (3) are equivalent. Here  $Q$  is a symmetric matrix and  $\text{Id}$  denotes the identity matrix. These relations attest to the importance of finding novel and efficient iterative procedures for solving equation (1).

In this paper, we first focus our attention on the *semi-smooth Newton method* for solving problem (1), which consists of specifying a particular generalized Jacobian of  $F$  at  $x$  in the problem of finding the zeroes of

$$F(x) := x^+ + Tx - b, \quad x \in \mathbb{R}^n. \tag{4}$$

Namely, starting at the point  $x^0 \in \mathbb{R}^n$ , the semi-smooth Newton iteration is defined by the following linear equation:

$$\left( P(x^k) + T \right) x^{k+1} = b, \quad k \in \mathbb{N}, \tag{5}$$

where

$$P(x) := \text{diag}(\text{sgn}(x^+)), \quad x \in \mathbb{R}^n. \tag{6}$$

The above iteration was proposed in [10], which was shown to be *globally* convergent to a solution of problem (1) under suitable assumptions. It has been extensively studied in the literature for solving generalizations of equation (1); see, for instance, [2, 5, 6, 11]. We emphasize that the global and linear convergence of iteration (5) has been proved only under restricted assumptions related to the norm of the matrices  $(T + P(x))^{-1}$  for all  $x \in \mathbb{R}^n$ , which come from the study of (5) as a contraction fixed point iteration. A promising and novel approach for establishing finite convergence of (5) was proposed in Theorem 3 of [9] under the assumption that the rows of the matrices  $(T + P(x))^{-1}$  for all  $x \in \mathbb{R}^n$  have a definite sign, that is, in every row the entries have the same sign. It is worth noting that the number of matrices  $P(x)$  with  $x \in \mathbb{R}^n$  in (6) is finite ( $2^n$  to be precise). Hence, if the semi-smooth Newton method (5) converges, this convergence will occur after finitely many steps. In the pursuit of weaker and verifiable sufficient conditions ensuring convergence of the sequence generated by (5), it was conjectured in [2] that iteration (5) converges after finitely many steps if  $T$  is symmetric and positive definite. In this paper, we show that this conjecture is false with a counterexample. However, interestingly, we show that this assumption is enough to guarantee the existence and uniqueness of solutions to problem (1). Moreover, the inverse of  $P(x) + T$  for all  $x \in \mathbb{R}^n$  always exists, and thus the semi-smooth Newton method (5) is well-defined. We will also show that although this method may cycle under this assumption, it can never cycle between only two points. In the second part of this paper, we propose two novel iterative processes inspired by the well-known Jacobi and Gauss-Seidel iterative methods for solving linear systems of equations. The main idea is to consider the Newtonian system (5) and apply a Jacobi or a Gauss-Seidel step at each iteration  $k$ . The main advantage of doing so is that the iteration is computed by solving a diagonal or a triangular linear system of equations, which is considerably simpler than solving the linear system in (5) to find  $x^{k+1}$ . Also, we are able to present new sufficient

conditions for the convergence of these two proposed methods, which are related to the classical diagonal dominance and Sassenfeld's criterion. The existence and uniqueness of the solution for equation (1) are also proved under both conditions. We then show with an example that the standard diagonal dominance is insufficient to ensure the existence of solutions to problem (1). Finally, numerical results show that the proposed methods are competitive in terms of CPU time if compared with the semi-smooth Newton method (5). The numerical illustration suggests that the proposed iterative methods become more efficient when the dimension is high, and the matrix is sparse. We finish the paper with an applied experiment with real data, where we solve and discuss the results of solving a piecewise linear equation raising from a discretization of the Boussinesq PDE [12] modeling a two-dimensional flow in a homogeneous phreatic aquifer.

## 1.1 Notations and preliminaries

Next, we quickly present some notations and facts used throughout the paper. We write  $\mathbb{N}$  for the nonnegative integers  $\{0, 1, 2, \dots\}$ . The canonical inner product in  $\mathbb{R}^n$  will be denoted by  $\langle \cdot, \cdot \rangle$  and the induced norm is  $\|\cdot\|$ . For  $x \in \mathbb{R}^n$ ,  $\text{sgn}(x)$  will denote a vector with components equal to 1, 0 or  $-1$  depending on whether the corresponding component of the vector  $x$  is positive, zero or negative, respectively. If  $a \in \mathbb{R}$  and  $x \in \mathbb{R}^n$ , then denote  $a^+ := \max\{a, 0\}$ ,  $a^- := \max\{-a, 0\}$  and  $x^+$  and  $x^-$  the vectors with  $i$ -th component equal to  $x_i^+$  and  $x_i^-$ , respectively,  $i = 1, \dots, n$ . Note that  $x^+$  and  $x^-$  are the projections of  $x$  onto the cones  $\mathbb{R}_+^n$  and  $\mathbb{R}_-^n$ , respectively. The matrix  $\text{Id} \in \mathbb{R}^{n \times n}$  denotes the identity matrix. If  $x \in \mathbb{R}^n$  then  $\text{diag}(x) \in \mathbb{R}^{n \times n}$  will denote a diagonal matrix with  $(i, i)$ -th entry equal to  $x_i$ ,  $i = 1, \dots, n$ . Denote  $\|M\| := \max\{\|Mx\| : x \in \mathbb{R}^n, \|x\| = 1\}$  for any  $M \in \mathbb{R}^{n \times n}$ . The following result is well-known and will be needed in the sequel:

**Theorem 1.1** (Contraction mapping principle [13], Thm. 8.2.2, page 153). *Let  $\Phi : \mathbb{R}^n \rightarrow \mathbb{R}^n$ . Suppose that there exists  $\lambda \in [0, 1)$  such that  $\|\Phi(y) - \Phi(x)\| \leq \lambda\|y - x\|$ , for all  $x, y \in \mathbb{R}^n$ . Then, there exists a unique  $\bar{x} \in \mathbb{R}^n$  such that  $\Phi(\bar{x}) = \bar{x}$ .*

## 2 The semi-smooth Newton method

In this section, we present and analyze the convergence of the semi-smooth Newton method given by iteration (5) for solving problem (1). In [9, 10], it was shown that the condition  $\text{sgn}((x^k)^+) = \text{sgn}((x^{k+1})^+)$  is sufficient to declare that  $x^{k+1}$  is a solution of equation (1). We now show, in fact, that a component-wise version of this stopping criterion holds:

**Proposition 2.1** (Component-wise stopping criterion). *Assume that the sequence  $(x^k)_{k \in \mathbb{N}}$  generated by method (5) is well defined. If  $\text{sgn}((x_i^{k+1})^+) = \text{sgn}((x_i^k)^+)$  for some  $i \in \{1, \dots, n\}$  and some  $k$ , then  $F_i(x^{k+1}) = 0$ , where  $F$  is defined in (4).*

*Proof.* By the definition of  $F$  and  $x^{k+1}$ , we have

$$\begin{aligned} F(x^{k+1}) &= (x^{k+1})^+ + Tx^{k+1} - b \\ &= (P(x^{k+1}) + T)x^{k+1} - b \\ &= (P(x^{k+1}) + T)x^{k+1} - (P(x^k) + T)x^{k+1} \\ &= (P(x^{k+1}) - P(x^k))x^{k+1}. \end{aligned}$$

Hence,

$$F_i(x^{k+1}) = \left( \operatorname{sgn}((x_i^{k+1})^+) - \operatorname{sgn}((x_i^k)^+) \right) x_i^{k+1}, \quad (7)$$

for any  $i \in \{1, \dots, n\}$ , which implies the desired result.  $\square$

In our study, a crucial role will be played by diagonal dominance. Let us start by showing that in the most extreme case of diagonal dominance, namely when the matrix is diagonal, one can list all solutions of the equation, and iteration (5) finds a solution in at most two steps.

**Proposition 2.2** (Finite convergence for the diagonal case). *Let  $b \in \mathbb{R}^n$  and  $T \in \mathbb{R}^{n \times n}$  be a diagonal matrix with entries  $T = (t_{ii})$ ,  $i = 1, \dots, n$  such that  $t_{ii} \notin \{0, -1\}$  for all  $i$ . Equation (1) has no solutions if, and only if,  $t_{ii} \in (-1, 0)$  and  $b_i < 0$  for some  $i$ . If a solution of (1) exists, then (5) converges in at most two iterations to one of the solutions. In this case, the number of solutions of problem (1) is given by  $2^r$ , where  $r$  is the number of indexes  $i$  such that  $b_i > 0$  and  $t_{ii} \in (-1, 0)$ .*

*Proof.* Let  $x^0 \in \mathbb{R}^n$  be any starting point. By definition, we have

$$x^1 = (P(x^0) + T)^{-1}b.$$

Since  $T$  is a diagonal matrix, we have the following component-wise expression for  $x^1$ :

$$x_i^1 = ((P(x^0) + T)^{-1}b)_i = \begin{cases} \frac{b_i}{t_{ii}}, & x_i^0 \leq 0, \\ \frac{b_i}{1+t_{ii}}, & x_i^0 > 0, \end{cases} \quad (8)$$

for  $i = 1, \dots, n$ . For a fixed  $i$ , if  $t_{ii} > 0$ , we get by (8) that  $\operatorname{sgn}((x_i^1)^+) = \operatorname{sgn}(b_i^+)$  and since  $b$  is fixed we deduce that  $\operatorname{sgn}((x_i^2)^+) = \operatorname{sgn}((x_i^1)^+)$ . Similarly, if  $t_{ii} < -1$ , we conclude that  $\operatorname{sgn}((x_i^1)^+) = 1 - \operatorname{sgn}(b_i^+)$ . Hence,  $\operatorname{sgn}((x_i^2)^+) = \operatorname{sgn}((x_i^1)^+)$ . Thus, when there is no  $i$  such that  $t_{ii} \in (-1, 0)$ , by Proposition 2.1 we deduce that (5) converges in two steps. In particular one can check that in this case the solution is unique with  $i$ -th component equals to

$$\begin{cases} \frac{b_i}{t_{ii}}, & b_i \leq 0, \\ \frac{b_i}{1+t_{ii}}, & b_i > 0, \end{cases}$$

when  $t_{ii} > 0$ , and

$$\begin{cases} \frac{b_i}{1+t_{ii}}, & b_i \leq 0, \\ \frac{b_i}{t_{ii}}, & b_i > 0, \end{cases}$$

when  $t_{ii} < -1$ .

Now, it directly follows from equation (1) that there is no solution if  $b_i < 0$  for some  $i$  such that  $t_{ii} \in (-1, 0)$ . If, however,  $b_i \geq 0$  for such  $i$ , the solutions for each component-wise equation are given by  $\frac{b_i}{t_{ii}}$  and  $\frac{b_i}{1+t_{ii}}$ , amounting to the desired formula for the number of solutions. Therefore, assuming that the problem has a solution, we conclude that  $b_i \geq 0$  for all  $i$  such that  $t_{ii} \in (-1, 0)$ . By (8), it is easy to see that in this case, we have  $\operatorname{sgn}((x_i^1)^+) = \operatorname{sgn}((x_i^0)^+)$ . Using Proposition 2.1 and a similar computation done previously, we conclude that the method converges to a solution in at most two steps.  $\square$

An auxiliary result in our analysis follows next.

**Proposition 2.3.** *Let  $x, y \in \mathbb{R}^n$ . Then*

$$i) (y_i - x_i)(\operatorname{sgn}(y_i^+)x_i - \operatorname{sgn}(x_i^+)y_i) \leq 0, \quad \forall i = 1, \dots, n;$$

$$ii) (y_i - x_i)(\operatorname{sgn}(x_i^+)x_i - \operatorname{sgn}(y_i^+)y_i) \leq 0, \quad \forall i = 1, \dots, n.$$

In particular we have that  $(y - x)^T(P(y)x - P(x)y) \leq 0$  and  $(y - x)^T(P(x)x - P(y)y) \leq 0$ .

*Proof.* We only have three different cases to be analyzed:

$$1) \text{ if } \operatorname{sgn}(y_i^+) = \operatorname{sgn}(x_i^+) = s, \text{ then } (y_i - x_i)(\operatorname{sgn}(y_i^+)x_i - \operatorname{sgn}(x_i^+)y_i) = -s(y_i - x_i)^2 \leq 0;$$

$$2) \text{ if } \operatorname{sgn}(y_i^+) = 1 \text{ and } \operatorname{sgn}(x_i^+) = 0, \text{ then } (y_i - x_i)(\operatorname{sgn}(y_i^+)x_i - \operatorname{sgn}(x_i^+)y_i) = x_i(y_i - x_i) \leq 0;$$

$$3) \text{ if } \operatorname{sgn}(y_i^+) = 0 \text{ and } \operatorname{sgn}(x_i^+) = 1, \text{ then } (y_i - x_i)(\operatorname{sgn}(y_i^+)x_i - \operatorname{sgn}(x_i^+)y_i) = -y_i(y_i - x_i) \leq 0.$$

Therefore, as a consequence we obtain that

$$(y - x)^T(P(x)y - P(y)x) = \sum_{i=1}^n (y_i - x_i)(\operatorname{sgn}(y_i^+)x_i - \operatorname{sgn}(x_i^+)y_i) \leq 0.$$

The second statement follows by a similar computation.  $\square$

Until the end of this section, we assume that the matrix  $T$  is symmetric and positive definite. This assumption has been considered before in [2] where the authors conjectured the global and finite convergence of the method under this assumption. To fully address this conjecture, we begin with a result about the existence and uniqueness of the solution of equation (1) under this assumption.

**Proposition 2.4** (Existence and uniqueness of solutions). *If  $T$  is a symmetric and positive definite  $n \times n$  matrix and  $b \in \mathbb{R}^n$ , then problem (1) has one and only one solution.*

*Proof.* Note first that the matrix  $T + \operatorname{Id}$  is also symmetric and positive definite. Hence,  $(T + \operatorname{Id})^{-1}$  exists. For any symmetric matrix  $M$  we denote  $\lambda_1(M) \leq \dots \leq \lambda_n(M)$  the ordered eigenvalues of  $M$ . The matrix  $T + \operatorname{Id}$  also satisfies  $\lambda_i(T + \operatorname{Id}) = 1 + \lambda_i(T)$ ,  $i = 1, \dots, n$ . Since  $\lambda_1(T) > 0$ , we deduce that

$$\begin{aligned} \|(T + \operatorname{Id})^{-1}\|_2 &= \lambda_n((T + \operatorname{Id})^{-1}) \\ &= \frac{1}{\lambda_1(T + \operatorname{Id})} \\ &= \frac{1}{\lambda_1(T) + 1} \\ &< 1. \end{aligned}$$

Defining the function  $\Phi(x) := (T + \operatorname{Id})^{-1}(b - x^-)$  and using the decomposition  $x = x^+ - x^-$  we can easily show that the solutions of problem (1) coincide with the fixed points of  $\Phi$ . Moreover,  $\Phi$  is a contraction and then has a unique fixed point. Indeed,

$$\begin{aligned} \|\Phi(x) - \Phi(y)\|_2 &= \|(T + \operatorname{Id})^{-1}(y^- - x^-)\|_2 \\ &\leq \|(T + \operatorname{Id})^{-1}\|_2 \|y^- - x^-\|_2 \\ &\leq \|(T + \operatorname{Id})^{-1}\|_2 \|x - y\|_2 \\ &= \frac{1}{\lambda_1(T) + 1} \|x - y\|_2, \end{aligned}$$

using that  $x^-$  and  $y^-$ , the projections of  $x$  and  $y$  onto  $\mathbb{R}^n$  (a closed and convex set), are non-expansive in the last inequality. Thus, the result follows from Theorem 1.1.  $\square$

**Remark 1** (Another proof of the uniqueness of solutions). *The uniqueness of the solution of equation (1) when  $T$  is positive definite showed in Theorem 2.4 can also be deduced using the fact*

$$(y - x)^T(P(x)x - P(y)y) \leq 0, \quad (9)$$

as seen in Proposition 2.3. Indeed, if there exist two solutions  $x$  and  $y$  of problem (1), we have by definition  $x^+ + Tx = b$ ,  $y^+ + Ty = b$ . Subtracting those equations and multiplying the resulting equation by  $(y - x)^T$ , we get that

$$(y - x)^T T(y - x) = (y - x)^T(x^+ - y^+) \leq 0,$$

where the inequality follows from (9) noting that  $P(x)x = x^+$  and  $P(y)y = y^+$ . Thus, when  $T$  is positive definite it must hold that  $x = y$ .

Due to the nature of iteration (5), we have that the sequence  $(x^k)_{k \in \mathbb{N}}$  has only a finite number of different elements. This happens since the set  $S := \{(P(x) + T)^{-1}b ; x \in \mathbb{R}^n\}$  has at most  $2^n$  different elements and  $(x^k)_{k \in \mathbb{N}} \subseteq S$ . The conclusion of this observation is that we have only two possible outcomes: the method converges in a finite number of steps, or it cycles. Note also that when  $\bar{x}$  solves equation (1), we have  $\bar{x} = (P(\bar{x}) + T)^{-1}b$ . Hence, by (5),  $x^{k+1} = \bar{x}$  holds whenever  $P(x^k) = P(\bar{x})$ . Thus, the problem amounts to finding a point in the same orthant as a solution. Let us first show that the method can only cycle among three or more points.

**Theorem 2.1** (Newton does not cycle between two points). *Let  $T \in \mathbb{R}^{n \times n}$  be a symmetric and positive definite matrix and  $b \in \mathbb{R}^n$ , then the sequence generated by the semi-smooth Newton method (5) does not cycle between two points.*

*Proof.* If the sequence generated by iteration (5) cycles between two points  $x$  and  $y$ , then

$$(P(x) + T)y = b,$$

$$(P(y) + T)x = b.$$

So, we have that  $T(y - x) = P(y)x - P(x)y$ . Multiplying by  $(y - x)^T$ , we obtain from Proposition 2.3 that

$$(y - x)^T T(y - x) = (y - x)^T(P(y)x - P(x)y) \leq 0,$$

which implies  $x = y$ . □

When  $n \leq 2$ , let us show that the sequence generated by method (5) in fact does not cycle.

**Theorem 2.2** (Finite convergence for low dimensions). *Let  $b \in \mathbb{R}^n$  and  $T \in \mathbb{R}^{n \times n}$  be a symmetric and positive definite matrix. Then, iteration (5) has global and finite convergence for  $n = 1, 2$ .*

*Proof.* The case  $n = 1$  is a direct consequence of Proposition 2.2. When  $n = 2$ , note that the sequence of Newton iterates has at most four different elements, however, no cycle of size four is possible, as these points are necessarily in four different quadrants of  $\mathbb{R}^2$ , coinciding in sign with the solution (which necessarily exists due to Proposition 2.4), which implies convergence.

Now, suppose that there is a cycle among three different points  $x, y, z$ , different from  $\bar{x}$ , the unique solution. Clearly, these four points must lie in different quadrants of  $\mathbb{R}^2$ . Without loss of

generality, let us assume that  $x$  and  $y$  lie in opposite quadrants, in the sense that  $P(x) + P(y) = \text{Id}$ , and that the points satisfy the following equations:

$$(P(x) + T)y = b,$$

$$(P(y) + T)z = b,$$

$$(P(z) + T)x = b.$$

We have that  $T(y - x) = P(z)x - P(x)y$ , and therefore, multiplying by  $(y - x)^T$ , we obtain

$$\begin{aligned} (y - x)^T T(y - x) &= (y - x)^T (P(z)x - P(x)y) \\ &= (y_1 - x_1)(\text{sgn}(z_1^+)x_1 - \text{sgn}(x_1^+)y_1) + (y_2 - x_2)(\text{sgn}(z_2^+)x_2 - \text{sgn}(x_2^+)y_2). \end{aligned}$$

Since  $z$  is not in the opposite quadrant of  $x$  we only have two options,  $\text{sgn}(z_1^+) = \text{sgn}(x_1^+)$  and  $\text{sgn}(z_2^+) \neq \text{sgn}(x_2^+)$ , or  $\text{sgn}(z_1^+) \neq \text{sgn}(x_1^+)$  and  $\text{sgn}(z_2^+) = \text{sgn}(x_2^+)$ . In both cases it can be checked that  $(y - x)^T T(y - x) \leq 0$ , which leads to a contradiction. The result now follows from Theorem 2.1.  $\square$

We ran extensive numerical experiments in our pursuit of understanding the finite convergence of the semi-smooth Newton method (5). For symmetric and positive definite matrices  $T \in \mathbb{R}^{n \times n}$  (1000 randomly generated problems for each dimension  $n$ ), we recorded the number of iterations that (5) needed to converge at each dimension. The results are shown below in Table 1, where the percentage of problems solved for several different problem dimensions and the corresponding iterations are presented.

$n/\%$	1 iter.	2 iter.	3 iter.	4 iter.	5 iter.
4	7.2	49.1	35.9	7	0.8
8	0.6	37.7	48.8	11.4	1.5
16	0	16.6	63.1	19.5	0.8
32	0	6	69	24.1	0.9
64	0	1.6	68.5	29.4	0.5
128	0	0.3	60.1	39.1	0.5
256	0	0	57.3	41.9	0.8
512	0	0	50.9	48.8	0.3
1024	0	0	43	56.9	0.1
2048	0	0	36.6	63.2	0.2
4096	0	0	30.6	69.2	0.2

Table 1: Percentage of positive definite problems solved at each amount of iterations (iter.) for different dimensions ( $n$ ).

We observe that the higher the dimension, the more iterations are needed. Surprisingly, the method performs exceptionally well as the required number of iterations grows very slowly (always capped by 5) with respect to the growth of the dimension  $n$ . Indeed, for any  $n \leq 4096$ , no problem needed more than five iterations of (5) to be solved.

In the tests illustrated in Table 2, the matrix  $T$  is symmetric, positive definite, and “almost” diagonal, *i.e.*, the elements of the diagonal are much greater compared with the off-diagonal entries. The diagonal entries are of the order of thousands, while the off-diagonal elements are smaller than

$n/\%$	1 iter.	2 iter.	3 iter.
4	0	98.9	1.1
8	0	100	0
16	0	100	0
32	0	100	0
64	0	100	0
128	0	99.9	0.01
256	0	98.9	1.1
512	0	95	5
1024	0	86.6	13.4
2048	0	74.8	25.2
4096	0	54.9	45.1

Table 2: Numerical experiments for “almost” diagonal positive definite matrices.

one. Here, as suggested by Proposition 2.2, no problem required more than three iterations to reach the solution.

Although the numerical tests based on random data suggest that the semi-smooth Newton method does not cycle, we were able to find a particular example of problem (1) that shows that the Newton iteration (5) may cycle among three points in  $\mathbb{R}^3$  even when  $T$  is symmetric and positive definite. This gives us a counterexample to the conjecture raised in [2] on the global and finite convergence of the Newton iteration (5) under this assumption.

**Example 1** (Counterexample on the finite convergence). *The semi-smooth Newton method (5) fails to converge in the case  $n = 3$  with the following symmetric and positive definite matrix*

$$T = \frac{1}{100} \begin{pmatrix} 32 & -26 & 21 \\ -26 & 33 & -23 \\ 21 & -23 & 17 \end{pmatrix} \quad \text{and} \quad b = \frac{1}{100} \begin{pmatrix} 18 \\ -48 \\ 30 \end{pmatrix}. \quad (10)$$

*It can be easily checked that the following points  $x, y$  and  $z$  conform a cycle of the method.*

$$x = \begin{pmatrix} \frac{319}{1435} \\ -\frac{1849}{6379} \\ \frac{190}{1191} \end{pmatrix}, \quad y = \begin{pmatrix} -\frac{527}{2978} \\ -\frac{1490}{923} \\ -\frac{81}{2777} \end{pmatrix}, \quad z = \begin{pmatrix} -\frac{306}{95} \\ \frac{18}{95} \\ 6 \end{pmatrix}.$$

*Indeed, those points satisfy the cycle equations*

$$(P(x) + T)y = b,$$

$$(P(y) + T)z = b,$$

$$(P(z) + T)x = b,$$

*which proves the statement.*

### 3 Jacobi-Newton and Gauss-Seidel-Newton methods

Based on the well-known Jacobi and Gauss-Seidel methods for solving linear systems, we define and analyze two novel methods, which we call Jacobi-Newton and Gauss-Seidel-Newton methods, for solving problem (1). We start with two definitions related to classical diagonal dominance and Sassenfeld's criterion.

**Definition 3.1** (Strong diagonal dominance). *Let  $T = (t_{ij}) \in \mathbb{R}^{n \times n}$ . We say that  $T$  is strongly diagonal dominant if*

$$\frac{1}{|t_{ii}|} \left( 1 + \sum_{\substack{j=1 \\ j \neq i}}^n |t_{ij}| \right) < 1, \quad \forall i = 1, \dots, n.$$

Note that if  $T$  is strongly diagonal dominant, then  $T$  is diagonal dominant. However, we need this stronger condition to ensure the global convergence of the Jacobi-Newton method, which will be presented later in (13).

We now introduce a weaker condition for  $T$ , which is a variation of the classical Sassenfeld's condition.

**Definition 3.2** (Strong Sassenfeld's condition). *Let  $T = (t_{ij}) \in \mathbb{R}^{n \times n}$ . Define  $\beta_i$  and  $\beta$  as follows*

$$\beta_1 := \frac{1}{|t_{11}|} \left( 1 + \sum_{j=2}^n |t_{1j}| \right), \quad (11)$$

$$\beta_i := \frac{1}{|t_{ii}|} \left( \sum_{j=1}^{i-1} |t_{ij}| \beta_j + \sum_{j=i+1}^n |t_{ij}| + 1 \right), \quad \forall i = 2, \dots, n, \quad (12)$$

and

$$\beta := \max_{i=1, \dots, n} \beta_i.$$

We say that  $T$  satisfies the strong Sassenfeld's condition if  $\beta < 1$ .

It is easy to see that if  $T$  is strongly diagonal dominant, then  $T$  satisfies the strong Sassenfeld's condition. However, the converse implication is not true in general. Now we prove the existence and uniqueness of solutions for problem (1) under the strong Sassenfeld's condition. This criterion will also be used later to prove the convergence of the Gauss-Seidel-Newton method, which we will introduce.

**Theorem 3.1** (Existence and uniqueness of solutions under strong Sassenfeld's condition). *Let  $b \in \mathbb{R}^n$  and  $T \in \mathbb{R}^{n \times n}$ . If  $T$  satisfies the strong Sassenfeld's condition, then equation (1) has one and only one solution.*

*Proof.* To prove the existence and uniqueness of the solution, we will use the contraction mapping principle (Theorem 1.1). Let us first decompose the matrix  $T$  as a sum of  $L + D + U$  where  $D$  is the diagonal of  $T$  and  $L$  and  $U$  are the strictly lower and upper parts of  $T$ , respectively. We first define the mapping  $\psi : \mathbb{R}^n \rightarrow \mathbb{R}^n$  such that

$$\psi(x) := (D + L)^{-1}(-Ux + b - x^+), \quad \forall x \in \mathbb{R}^n.$$

Then, we prove that the fixed points of  $\psi$  are solutions of problem (1) and also that it is a contraction. Indeed, note that  $x = \psi(x)$  is equivalent to  $(D+L)x = -Ux+b-x^+$ , or  $(D+L+U)x+x^+ = b$ . Hence,  $x$  is solution of equation (1) given that  $T = D + L + U$ .

On the other hand, to prove the contraction property of  $\psi$  consider any  $x, y \in \mathbb{R}^n$  and define  $u := \psi(x)$  and  $w := \psi(y)$ . Note that

$$u - w = \psi(x) - \psi(y) = (D + L)^{-1}(-U(x - y) + y^+ - x^+),$$

which is equivalent to

$$(D + L)(u - w) = -U(x - y) + y^+ - x^+.$$

Hence,

$$u_1 - w_1 = \frac{1}{t_{11}} \left( - \sum_{j=2}^n t_{1j}(x_j - y_j) + y_1^+ - x_1^+ \right),$$

and after taking the absolute value in both sides and using its triangular inequality property, we have

$$\begin{aligned} |u_1 - w_1| &\leq \frac{1}{|t_{11}|} \left( \sum_{j=2}^n |t_{1j}| |x_j - y_j| + |x_1^+ - y_1^+| \right) \\ &\leq \frac{1}{|t_{11}|} \left( \sum_{j=2}^n |t_{1j}| |x_j - y_j| + |x_1 - y_1| \right) \\ &\leq \frac{1}{|t_{11}|} \left( \sum_{j=2}^n |t_{1j}| + 1 \right) \|x - y\|_\infty \\ &= \beta_1 \|x - y\|_\infty. \end{aligned}$$

The second inequality follows from the nonexpansiveness of the projections and the last equality by the definition of  $\beta_1$  given in (11). By induction let us assume that  $|u_i - w_i| \leq \beta_i \|x - y\|_\infty$  for all  $i = 1, \dots, p-1$ , and we will show that  $|u_p - w_p| \leq \beta_p \|x - y\|_\infty$ . Analogously,

$$u_p - w_p = \frac{1}{t_{pp}} \left( - \sum_{j=1}^{p-1} t_{pj}(u_j - w_j) - \sum_{j=p+1}^n t_{pj}(x_j - y_j) + y_p^+ - x_p^+ \right),$$

which implies that

$$\begin{aligned} |u_p - w_p| &\leq \frac{1}{|t_{pp}|} \left( \sum_{j=1}^{p-1} |t_{pj}| \beta_j + 1 + \sum_{j=p+1}^n |t_{pj}| \right) \|x - y\|_\infty \\ &= \beta_p \|x - y\|_\infty, \end{aligned}$$

where we used the induction assumption and the nonexpansiveness of the projection, together with the definition of  $\beta_p$  given in (12) in the last equality. Now, we are ready to prove the contraction property of  $\psi$  by observing that for all  $x, y \in \mathbb{R}^n$ , we have

$$\begin{aligned} \|\psi(x) - \psi(y)\|_\infty &= \|u - w\|_\infty \\ &= \max_{i=1, \dots, n} |u_i - w_i| \\ &\leq \beta \|x - y\|_\infty. \end{aligned}$$

Thus, since  $\beta < 1$ , we have that  $\psi$  is a contraction mapping and therefore has a unique fixed point.  $\square$

Clearly, if  $T$  is strongly diagonal dominant, the result above is also true, *i.e.*, problem (1) has one and only one solution. A natural question that arises here is if this assumption can be relaxed to the classical diagonal dominance and if the existence and uniqueness of solutions would still hold for equation (1). Unfortunately, that is not true in general, as is shown in the following example.

**Example 2** (Diagonal dominance is not sufficient for existence of solutions). *Let  $b \in \mathbb{R}^n$  and  $T$  a diagonal dominant matrix. Then, problem (1) may fail to have solutions. Indeed, let us consider the following data*

$$T = \frac{1}{100} \begin{pmatrix} -26 & 16 \\ 23 & -33 \end{pmatrix}, \quad b = \frac{1}{100} \begin{pmatrix} -12 \\ 12 \end{pmatrix}.$$

*Note that  $T$  is diagonal dominant and the points  $x, y$  and  $z, w$  conform to two different cycles for the Newton iteration (5) where*

$$x = \frac{1}{2295} \begin{pmatrix} -498 \\ 582 \end{pmatrix}, \quad y = \frac{1}{1055} \begin{pmatrix} 498 \\ 18 \end{pmatrix},$$

*and*

$$z = \frac{1}{245} \begin{pmatrix} 102 \\ -18 \end{pmatrix}, \quad w = -\frac{1}{1405} \begin{pmatrix} 102 \\ 582 \end{pmatrix}.$$

*Namely, the cycle equations are satisfied:  $(P(x) + T)y = b$ ,  $(P(y) + T)x = b$  and  $(P(z) + T)w = b$ ,  $(P(w) + T)z = b$ . Thus, the method has two cycles of order two and since the points are in different quadrants we have that the equation has no solution (see the discussion before Theorem 2.1).*

To define the Jacobi-Newton iteration for solving problem (1), let us recall the decomposition of the matrix  $T$  as a sum of  $L + D + U$  where  $D$  is the diagonal of  $T$  and  $L$  and  $U$  are the strictly lower and upper parts of  $T$ , respectively. The iteration is defined as follows: Given  $x^0 \in \mathbb{R}^n$ ,

$$x^{k+1} := -(P(x^k) + D)^{-1}(L + U)x^k + (P(x^k) + D)^{-1}b, \quad k \in \mathbb{N}, \quad (13)$$

or equivalently,  $x^{k+1}$  is the solution of the diagonal system

$$(P(x^k) + D)x^{k+1} = -(L + U)x^k + b. \quad (14)$$

First, we prove that if the sequence generated by the Jacobi-Newton iteration (13) converges to  $\bar{x} \in \mathbb{R}^n$ , then  $\bar{x}$  is a solution of problem (1).

**Proposition 3.1** (Limit points of the Jacobi-Newton iteration are solutions). *If the sequence  $(x^k)_{k \in \mathbb{N}}$  generated by (13) converges to  $\bar{x}$ , then  $\bar{x}$  solves problem (1).*

*Proof.* First, we know that  $F(x) = x^+ + Tx - b$  is continuous since  $x^+$  is a continuous piecewise linear function and  $Tx - b$  is linear and therefore continuous. Rewriting  $F(x^{k+1})$  and using (14), we have

$$\begin{aligned} F(x^{k+1}) &= (P(x^{k+1}) + T)x^{k+1} - b, \\ &= (P(x^{k+1}) + T)x^{k+1} - (P(x^k) + D)x^{k+1} - (L + U)x^k, \\ &= (P(x^{k+1}) + D + L + U)x^{k+1} - (P(x^k) + D)x^{k+1} - (L + U)x^k, \\ &= (P(x^{k+1}) - P(x^k))x^{k+1} + (L + U)(x^{k+1} - x^k), \end{aligned}$$

using (14) in the last equality. Now, the above equality implies that

$$\|F(x^{k+1})\| \leq \|(P(x^{k+1}) - P(x^k))x^{k+1}\| + \|L + U\| \|x^{k+1} - x^k\|.$$

Hence,

$$\begin{aligned} \lim_{k \rightarrow \infty} \|F(x^{k+1})\| &\leq \lim_{k \rightarrow \infty} \|(P(x^{k+1}) - P(x^k))x^{k+1}\| + \lim_{k \rightarrow \infty} \|L + U\| \|x^{k+1} - x^k\| \\ &= \lim_{k \rightarrow \infty} \|(P(x^{k+1}) - P(x^k))x^{k+1}\|. \end{aligned}$$

If  $\bar{x}$  does not have any component equal to 0, then the result is trivial since the sign becomes constant for a certain index of the sequence. On the other hand, if  $\bar{x}_i = 0$  for some  $i$ , we have by the continuity of  $F$  that

$$|F_i(\bar{x})| = \lim_{k \rightarrow \infty} |F_i(x^{k+1})| \leq \lim_{k \rightarrow \infty} |(\text{sgn}((x_i^{k+1})^+) - \text{sgn}((x_i^k)^+))x_i^{k+1}| = 0,$$

where the last equality follows from the fact that the  $i$ -th component of  $x^{k+1}$  goes to zero and  $\text{sgn}((x_i^{k+1})^+) - \text{sgn}((x_i^k)^+) \in \{-1, 0, 1\}$ . Thus, all component of  $F(\bar{x})$  are zeroes, which implies the result.  $\square$

Next, we show that  $T$  being strongly diagonal dominant is a sufficient condition for the convergence of iteration (13).

**Theorem 3.2** (Global convergence of the Jacobi-Newton method). *Let  $b \in \mathbb{R}^n$  and  $T \in \mathbb{R}^{n \times n}$ . If  $T$  is strongly diagonal dominant, then the Jacobi-Newton method (13) globally converges to the unique solution of problem (1).*

*Proof.* Let  $\bar{x}$  be the unique solution of problem (1). Then,  $(P(\bar{x}) + D)\bar{x} = -(L + U)\bar{x} + b$ , which combined with (14) implies that

$$(P(x^k) + D)(x^{k+1} - \bar{x}) = -(L + U)(x^k - \bar{x}) + P(\bar{x})\bar{x} - P(x^k)\bar{x}.$$

Hence,

$$\begin{aligned} \|x^{k+1} - \bar{x}\|_\infty &= \|-(P(x^k) + D)^{-1}[(L + U)(x^k - \bar{x}) + (P(\bar{x}) - P(x^k))\bar{x}]\|_\infty \\ &= \max_{i=1, \dots, n} \left| \frac{1}{\text{sgn}((x_i^k)^+) + t_{ii}} \left( - \sum_{\substack{j=1 \\ j \neq i}}^n t_{ij}(x_j^k - \bar{x}_j) + \text{sgn}(\bar{x}_i^+) \bar{x}_i - \text{sgn}((x_i^k)^+) \bar{x}_i \right) \right| \\ &\leq \max_{i=1, \dots, n} \frac{1}{|\text{sgn}((x_i^k)^+) + t_{ii}|} \left( \sum_{\substack{j=1 \\ j \neq i}}^n |t_{ij}| |x_j^k - \bar{x}_j| + |\text{sgn}(\bar{x}_i^+) \bar{x}_i - \text{sgn}((x_i^k)^+) \bar{x}_i| \right). \end{aligned}$$

Note that

$$|\text{sgn}(\bar{x}_i^+) \bar{x}_i - \text{sgn}((x_i^k)^+) \bar{x}_i| = \begin{cases} 0, & \text{sgn}(\bar{x}_i^+) = \text{sgn}((x_i^k)^+), \\ |\bar{x}_i|, & \text{sgn}(\bar{x}_i^+) \neq \text{sgn}((x_i^k)^+). \end{cases}$$

Observe further that  $|\text{sgn}(\bar{x}_i^+) \bar{x}_i - \text{sgn}((x_i^k)^+) \bar{x}_i| = |\bar{x}_i| \leq |x_i^k - \bar{x}_i|$  when  $\text{sgn}(\bar{x}_i^+) \neq \text{sgn}((x_i^k)^+)$  and  $|\text{sgn}(\bar{x}_i^+) \bar{x}_i - \text{sgn}((x_i^k)^+) \bar{x}_i| = 0 \leq |x_i^k - \bar{x}_i|$  if  $\text{sgn}(\bar{x}_i^+) = \text{sgn}((x_i^k)^+)$ . Therefore,  $|\text{sgn}(\bar{x}_i^+) \bar{x}_i - \text{sgn}((x_i^k)^+) \bar{x}_i| \leq |x_i^k - \bar{x}_i|$ .

Then,

$$\begin{aligned}
\|x^{k+1} - \bar{x}\|_\infty &\leq \max_{i=1,\dots,n} \frac{1}{|\operatorname{sgn}((x_i^k)^+) + t_{ii}|} \left( \sum_{\substack{j=1 \\ j \neq i}}^n |t_{ij}| |x_j^k - \bar{x}_j| + |x_i^k - \bar{x}_i| \right) \\
&\leq \left( \max_{i=1,\dots,n} \frac{1}{|\operatorname{sgn}((x_i^k)^+) + t_{ii}|} \left( \sum_{\substack{j=1 \\ j \neq i}}^n |t_{ij}| + 1 \right) \right) \|x^k - \bar{x}\|_\infty \\
&\leq \left( \max_{i=1,\dots,n} \frac{1}{|t_{ii}|} \left( \sum_{\substack{j=1 \\ j \neq i}}^n |t_{ij}| + 1 \right) \right) \|x^k - \bar{x}\|_\infty.
\end{aligned}$$

The result now follows from the fact that  $T$  is strongly diagonal dominant.  $\square$

Now we define the Gauss-Seidel-Newton method as follows: Given  $x^0 \in \mathbb{R}^n$ ,

$$x^{k+1} := -(P(x^k) + D + L)^{-1}Ux^k + (P(x^k) + D + L)^{-1}b, \quad k \in \mathbb{N}, \quad (15)$$

or equivalently,  $x^{k+1}$  is the solution of the triangular linear system

$$(P(x^k) + D + L)x^{k+1} = -Ux^k + b. \quad (16)$$

**Proposition 3.2** (Limit points of the Gauss-Seidel-Newton method are solutions). *If the sequence  $(x^k)_{k \in \mathbb{N}}$  generated by (15) converges to  $\bar{x}$ , then  $\bar{x}$  solves problem (1).*

*Proof.* Let  $F(x) = x^+ + Tx - b$  and let us rewrite  $F(x^{k+1})$  and use (16) to obtain

$$\begin{aligned}
F(x^{k+1}) &= (P(x^{k+1}) + T)x^{k+1} - b, \\
&= (P(x^{k+1}) + T)x^{k+1} - (P(x^k) + D + L)x^{k+1} - Ux^k, \\
&= (P(x^{k+1}) + D + L + U)x^{k+1} - (P(x^k) + D + L)x^{k+1} - Ux^k, \\
&= (P(x^{k+1}) - P(x^k))x^{k+1} + U(x^{k+1} - x^k).
\end{aligned}$$

The result now follows analogously to Proposition 3.1.  $\square$

Now, let us show that the strong Sassenfeld's condition gives, besides the existence and uniqueness of a solution, a sufficient condition for global convergence of the Gauss-Seidel-Newton method to the solution.

**Theorem 3.3** (Global convergence under strong Sassenfeld's condition). *Let  $b \in \mathbb{R}^n$  and  $T \in \mathbb{R}^{n \times n}$ . If  $T$  satisfies the strong Sassenfeld's condition, then the Gauss-Seidel-Newton method (15) globally converges to the unique solution of problem (1).*

*Proof.* Let  $\bar{x}$  be the unique solution of equation (1). From (16), we have

$$(P(x^k) + D - L)(x^{k+1} - \bar{x}) = U(x^k - \bar{x}) + (P(\bar{x}) - P(x^k))\bar{x},$$

which implies

$$x_1^{k+1} - \bar{x}_1 = \frac{1}{\operatorname{sgn}((x_1^k)^+) + t_{11}} \left( - \sum_{j=2}^n t_{1j} (x_j^k - \bar{x}_j) + (\operatorname{sgn}(\bar{x}_1^+) - \operatorname{sgn}((x_1^k)^+)) \bar{x}_1 \right),$$

then

$$\begin{aligned} |x_1^{k+1} - \bar{x}_1| &\leq \frac{1}{|\operatorname{sgn}((x_1^k)^+) + t_{11}|} \left( \sum_{j=2}^n |t_{1j}| |x_j^k - \bar{x}_j| + |\operatorname{sgn}(\bar{x}_1^+) - \operatorname{sgn}((x_1^k)^+)| |\bar{x}_1| \right) \\ &\leq \frac{1}{|\operatorname{sgn}((x_1^k)^+) + t_{11}|} \left( \sum_{j=2}^n |t_{1j}| \|x^k - \bar{x}\|_\infty + |x_1^k - \bar{x}_1| \right) \\ &\leq \frac{1}{|\operatorname{sgn}((x_1^k)^+) + t_{11}|} \left( 1 + \sum_{j=2}^n |t_{1j}| \right) \|x^k - \bar{x}\|_\infty \\ &\leq \beta_1 \|x^k - \bar{x}\|_\infty. \end{aligned}$$

By induction let us assume that  $|x_i^{k+1} - \bar{x}_i| \leq \beta_i \|x^k - \bar{x}\|_\infty$  for  $i = 1, \dots, p-1$ , and we will show that  $|x_p^{k+1} - \bar{x}_p| \leq \beta_p \|x^k - \bar{x}\|_\infty$ . Now, similarly to the proof of Theorem 3.1, we have

$$\begin{aligned} x_p^{k+1} - \bar{x}_p &= \frac{1}{\operatorname{sgn}((x_p^k)^+) + t_{pp}} \left( - \sum_{j=1}^{p-1} t_{pj} (x_j^{k+1} - \bar{x}_j) - \sum_{j=p+1}^n t_{pj} (x_j^k - \bar{x}_j) \right. \\ &\quad \left. + (\operatorname{sgn}(\bar{x}_p^+) - \operatorname{sgn}((x_p^k)^+)) \bar{x}_p \right). \end{aligned}$$

Taking the absolute value in both sides, we get

$$\begin{aligned} |x_p^{k+1} - \bar{x}_p| &\leq \frac{1}{|\operatorname{sgn}((x_p^k)^+) + t_{pp}|} \left( \sum_{j=1}^{p-1} |t_{pj}| |x_j^{k+1} - \bar{x}_j| + \sum_{j=p+1}^n |t_{pj}| |x_j^k - \bar{x}_j| \right. \\ &\quad \left. + |\operatorname{sgn}(\bar{x}_p^+) - \operatorname{sgn}((x_p^k)^+)| |\bar{x}_p| \right), \\ &\leq \frac{1}{|\operatorname{sgn}((x_p^k)^+) + t_{pp}|} \left( \sum_{j=1}^{p-1} |t_{pj}| |x_j^{k+1} - \bar{x}_j| + \sum_{j=p+1}^n |t_{pj}| |x_j^k - \bar{x}_j| + |x_p^k - \bar{x}_p| \right), \\ &\leq \frac{1}{|\operatorname{sgn}((x_p^k)^+) + t_{pp}|} \left( \sum_{j=1}^{p-1} |t_{pj}| \beta_j + \sum_{j=p+1}^n |t_{pj}| + 1 \right) \|x^k - \bar{x}\|_\infty \\ &\leq \beta_p \|x^k - \bar{x}\|_\infty. \end{aligned}$$

Moreover,

$$\begin{aligned} \|x^{k+1} - \bar{x}\|_\infty &= \max_{i=1, \dots, n} |x_i^{k+1} - \bar{x}_i| \\ &\leq \max_{i=1, \dots, n} \beta_i \|x^k - \bar{x}\|_\infty \\ &= \beta \|x^k - \bar{x}\|_\infty, \end{aligned}$$

and the result follows from the fact that  $\beta < 1$ . □

## 4 Computational results

To analyze the three methods and see their differences in practice, we run several examples of problem (1) applying the Jacobi-Newton, Gauss-Seidel-Newton, and the semi-smooth Newton methods. All codes were implemented in Matlab 9.5.0.944444 (R2018b).

### 4.1 Numerical tests for randomly generated data

In this subsection, we work on two groups of problems for the numerical tests. In the first one, we used dense matrices with different dimensions  $n$  equal to 1000, 5000, and 10000. In the second group, we considered a matrix with sparse structure and dimensions 1000, 5000, and 10000. The experiments were run on a 2.3 GHz Intel(R) i5, 16Gb of RAM, and Windows 10 operating system. Next, we describe some details about the implementation.

(1) *Stopping criterion:* We fix the tolerance for the norm of  $F(x^k)$  in (4) as  $10^{-5}$ . This means that when the 2-norm of  $(x^k)^+ + Tx^k - b$  is less than or equal to  $10^{-5}$ , the execution of the algorithm is stopped, and  $x^k$  is returned as the solution. The maximum number of iterations was fixed at 1000, but no problem in our test reached it.

(2) *Generating random problems:* To construct the matrices for the first group of experiments, we used the Matlab routine *rand* to generate a random dense matrix with a predefined dimension with elements between  $-1$  and  $1$ . To achieve convergence of the different methods, we modified the matrices in order to ensure the validity of the strong diagonal dominance condition in Definition 3.1. To do so, we replaced the diagonal entry with  $1.001$  plus the sum of the off-diagonal elements in absolute value; this ensures the convergence of both Jacobi-Newton and Gauss-Seidel-Newton methods. Finally, for the second group of problems, we evoked *sprand*, a sparse random matrix generator of Matlab, to generate random sparse matrices with entries between  $-1$  and  $1$  and density  $0.3\%$ . The matrices were also similarly modified in order to ensure strong diagonal dominance.

(3) *Solving linear equations:* Each iteration of the semi-smooth Newton method requires solving the whole linear system (5). On the other hand, the Jacobi-Newton and Gauss-Seidel-Newton methods need to find the solution of a diagonal (14) and upper triangular (16) linear systems, respectively. We used the *backslash* command of Matlab for the dense case since this command uses the diagonal and triangular structures of the matrices generated by both methods (14) and (16). When solving Newton's linear system for the sparse case, we first use the *symamd* command of Matlab, which makes permutations of rows and columns using the minimum degree algorithm (see [14, 15]) to ensure that the LU-decomposition generates the lowest number of non-zero entries possible.

In order to show the results, we use performance profiles (see [16]). This technique measures and compares the robustness and efficiency of several methods applied to a set of problems using a common indicator, in our case, the CPU time. Let us start by presenting the set of tests with dense matrices.

#### 4.1.1 Piecewise linear equation with dense matrices

We took dense matrices with different dimensions in the first group of problems. We show in Figure 1 the results of the performance profiles in our experiments.

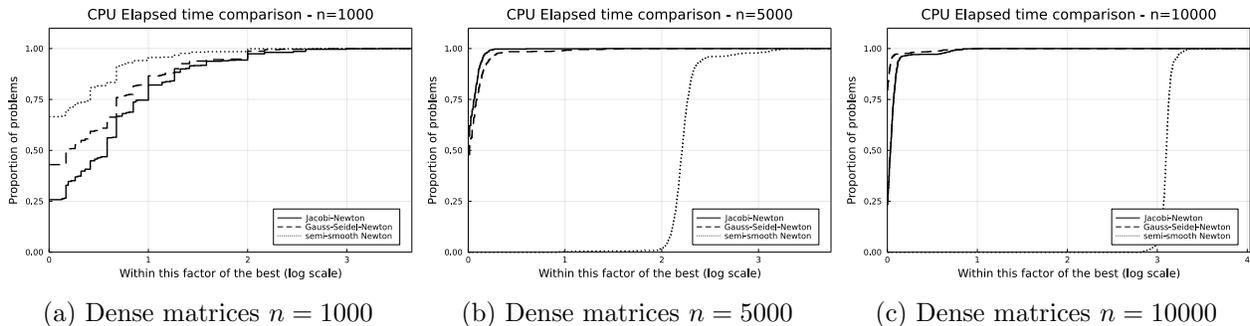


Figure 1: Performance profiles for dense matrices in  $\log_2$  scale.

In Figure 1, we compare the robustness and the efficiency of the three methods applied on a set of problems with dimensions 1000 (low), 5000 (mid), and 10000 (high), respectively. The number of problems was fixed at 850 for each dimension. We first see that for the three sets, every problem was solved by the three methods. Newton’s method was the most efficient in the low dimensional test, being the fastest method for circa 70% of the problems. When the dimension increases, Jacobi-Newton and Gauss-Seidel-Newton are much faster. This difference is accentuated in the highest dimension we tested, where also the Gauss-Seidel variant is now slightly better than Jacobi. In particular, Newton’s method took at least four times the time the other methods took for almost all mid-dimensional problems. At the same time, it was at least eight times slower for high-dimensional problems. Thus, in our tests, the simplicity of the linear system solved by Gauss-Seidel-Newton and Jacobi-Newton methods (triangular and diagonal, respectively) pays off in comparison with the Newton iteration for mid and high dimensions.

#### 4.1.2 Piecewise linear equation with sparse matrices

Our Jacobi and Gauss-Seidel variants of the Newton iterate were devised mainly with large and sparse problems in mind. In Figure 2, we can see the results of our numerical experiments for sparse matrices. Here we also generated 850 problems for each of the dimensions: 1000, 5000, and 10000.

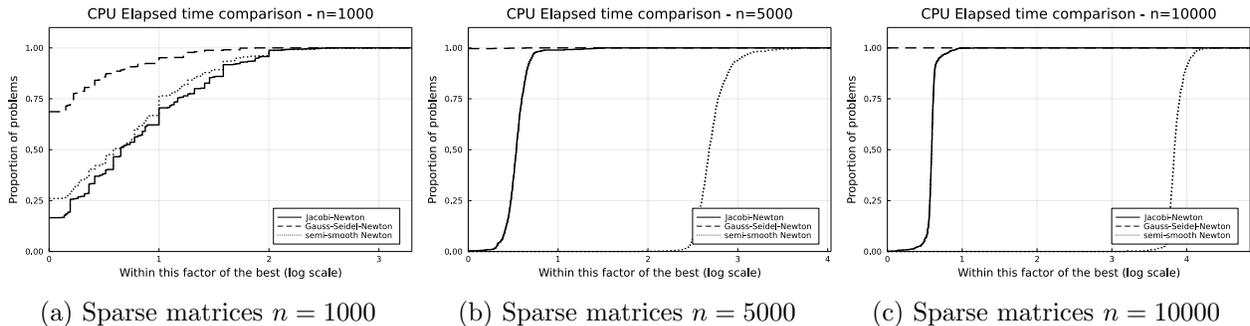


Figure 2: Performance profiles for sparse matrices in  $\log_2$  scale.

In these tests, again all problems were solved by all methods, but here, the superiority of the Gauss-Seidel-Newton iterate is already apparent in the low dimensional test, while Jacobi-Newton and Newton behave similarly. The superiority of Gauss-Seidel-Newton is more evident once the dimension increases, being the fastest method for almost all mid and high-dimensional problems.

At the same time, Newton becomes considerably slower than both methods. This behavior was already expected, and they attest that our Gauss-Seidel variant of Newton's method should be the method of choice for large and sparse problems.

## 4.2 Application on a discretization of the Boussinesq PDE

In order to test the proposed methods in solving a real model, we used them to solve an equation studied in [12]. The authors in [12] solve a piecewise linear equation in the form of problem (1) resulting from the discretization of a PDE, the Boussinesq equation, using the semi-smooth Newton method (see [17]), which models a two-dimensional flow of liquid water in a homogeneous phreatic aquifer during seven days. The Boussinesq equation models the water level in time, and a discretization of it results in an equation that can be solved by the semi-smooth Newton, the Jacobi-Newton, and the Gauss-Seidel-Newton methods. After the discretization using a square mesh of size  $2N + 1$  and the index  $l$  representing the respective day, the piecewise linear equation resulting has the following form:

$$\begin{aligned}
(h_{ij} + \eta_{ij}^{l+1})^+ - \left( \frac{\kappa \Delta t}{\epsilon} \frac{H_{i,j-\frac{1}{2}}^l}{\Delta y^2} \right) \eta_{ij-1}^{l+1} - \left( \frac{\kappa \Delta t}{\epsilon} \frac{H_{i-\frac{1}{2},j}^l}{\Delta x^2} \right) \eta_{i-1j}^{l+1} \\
+ \frac{\kappa \Delta t}{\epsilon} \left( \frac{H_{i+\frac{1}{2},j}^l + H_{i-\frac{1}{2},j}^l}{\Delta x^2} + \frac{H_{i,j+\frac{1}{2}}^l + H_{i,j-\frac{1}{2}}^l}{\Delta y^2} \right) \eta_{ij}^{l+1} - \left( \frac{\kappa \Delta t}{\epsilon} \frac{H_{i+\frac{1}{2},j}^l}{\Delta x^2} \right) \eta_{i+1j}^{l+1} \\
- \left( \frac{\kappa \Delta t}{\epsilon} \frac{H_{i,j+\frac{1}{2}}^l}{\Delta y^2} \right) \eta_{ij+1}^{l+1} = H_{ij}^l + \frac{\Delta t}{\epsilon} \phi_{ij}^l. \tag{17}
\end{aligned}$$

The parameters  $\epsilon := 0.4$  and  $\kappa := 1$  are the porosity and the hydraulic conductivity of the aquifer, respectively, while  $\Delta x := \frac{L}{N}$ ,  $\Delta y := \frac{L}{N}$ ,  $\Delta t := 86400\text{s}$  (1 day) are the stepsizes in the  $x$ -axis,  $y$ -axis, and the time step, respectively, where the aquifer is assumed to be a paraboloid of revolution with maximum radius  $L := 1000$  meters and depth of 10 meters. Water sinks from a pointwise source located at the bottom of the aquifer, which corresponds to the parameter  $\phi_{00}^l := -\frac{q}{\Delta x \Delta y}$ , for every  $l$ , where  $q := 10 \text{ m}^3/\text{s}$  is the volume of water sinking per second, with  $\phi_{ij}^l := 0$  for the remaining indexes  $i$  and  $j$ . The variables of the system (17) are  $\eta_{ij}^{l+1}$ ,  $l = 0, \dots, 6$  and  $i, j = -N, \dots, N$ , which represents the distance from the reference level 0 to the surface of water at the corresponding point in space and time, while  $h_{ij}$  is the given distance from the bottom of the aquifer to the reference level 0 at the correspondent point in space. Finally, the parameter  $H_{ij}^l := (h_{ij} + \eta_{ij}^l)^+$  is the total distance from the bottom of the aquifer to the surface level, which is obtained from the solution  $\eta_{ij}^l$  computed in the previous day, where the level of water at day zero sits at the reference level 0. The terms  $H_{i\pm\frac{1}{2},j}^l$  and  $H_{i,j\pm\frac{1}{2}}^l$  are defined as the averages of their nearest grid values. We refer to [12] for a more detailed description of this model. In order to write the system for the day  $l + 1$  in the format of (1), we considered the change of variables  $x_{ij} = h_{ij} + \eta_{ij}^{l+1}$ . The resulting piecewise linear system has a symmetric, positive semidefinite and block-tridiagonal matrix, specifically, the diagonal blocks are tridiagonal matrices, and the subdiagonal blocks are diagonal, so it has a pseudo pentadiagonal structure, meaning that we only need three vectors to save the entire matrix. This structure is exploited in order to compute the iterates of each method. For solving the system for each day  $l + 1 = 1, 2, \dots, 7$ , we used the Gauss-Seidel-Newton method since it showed better results for sparse large-scale matrices. Actually, Gauss-Seidel-Newton was about three times faster than

Jacobi-Newton in an initial test with  $N = 50$ . We used a tolerance of  $10^{-5}$  for the norm of  $F(x^k)$  in (4) to stop the iteration.

A significant detail in our implementation is the choice of the initial point for the method at each day. Having solved the system for the grid size  $N = 50$ , we used an interpolated version (completing the missing nodes with a mean using the nearest values) of that solution for smaller grid sizes, specifically  $N = 100$ . Then, we applied the same strategy for  $N = 200$  (using the levels for  $N = 100$ ). It is worth noting that this strategy helps the Gauss-Seidel-Newton method to converge faster, and it is different from the one used in [12] where the water level for the previous day is used as the initial point. Similarly, the solution for  $N = 50$  is obtained from the solution of  $N = 25$ , where the initial point from [12] was used for  $N = 25$ .

In table 3, we present the results in terms of time (in minutes) for each day, needed to solve each system (time), and the approximated volume of water in the phreatic aquifer at each day computed with the solution found for the equation. This serves as a measure of accuracy of the solution found as this volume can be computed from the model, that is, at day 0, the volume of water is equal to the volume of the aquifer times the porosity constant  $\kappa$ , which amounts to  $2,000,000\pi$  cubic meters. Thus, considering the constant flow of water  $q$ , we can predict the total volume of water at day 7 to be approximately 235,185.3 cubic meters, which is well approximated by the solution found with  $N = 200$ . Note that our computed volume coincides with the one computed in [12].

day\N	50		100		200	
-	time	volume	time	volume	time	volume
1	9.06	5,419,110.3	103.64	5,419,172.7	1561.03	5,419,182.2
2	8.27	4,555,110.2	93.80	4,555,172.7	1393.78	4,555,182.2
3	7.58	3,691,110.1	82.92	3,691,172.6	1273.69	3,691,182.1
4	7.25	2,827,109.9	72.98	2,827,172.6	1155.79	2,827,182.1
5	6.24	1,963,109.8	64.82	1,963,172.5	934.44	1,963,182.1
6	5.20	1,099,109.8	55.04	1,099,172.5	834.09	1,099,182.1
7	5.71	235,109.7	57.60	235,172.4	806.94	235,182.0

Table 3: Results for discretizations  $N = 50, 100, 200$  for the Gauss-Seidel-Newton method.

An important remark is the fact that the matrix defining the problem does not satisfy the sufficient condition for the global convergence of Gauss-Seidel-Newton (Theorem 3.3), which slows down the methods considerably in comparison with the standard Newton iterate. In particular our Newton implementation, as described previously, was about 20 times faster on average for this problem, requiring only 3 or 4 iterations. Nevertheless, both our proposed methods are still able to converge, showing robustness, which is the point of this numerical experiment. For comparison, at the cost of one Newtonian iteration for  $N = 50$ , we are able to compute approximately 1000 iterations of the Gauss-Seidel-Newton method or approximately 3000 iterations of the Jacobi-Newton method. For the Jacobi-Newton iterate one can easily make use of parallel computations to speed up the algorithm. Finally, in Figure 3, we draw a two-dimensional vertical cut of the approximated water levels found for each day with the Gauss-Seidel-Newton method for discretization  $N = 200$ .

## 5 Conclusions

In this paper, we considered iterative schemes for solving the piecewise linear equation  $x^+ + Tx = b$ , where  $x^+$  denotes projection onto the non-negative orthant. This problem appears in solving ab-

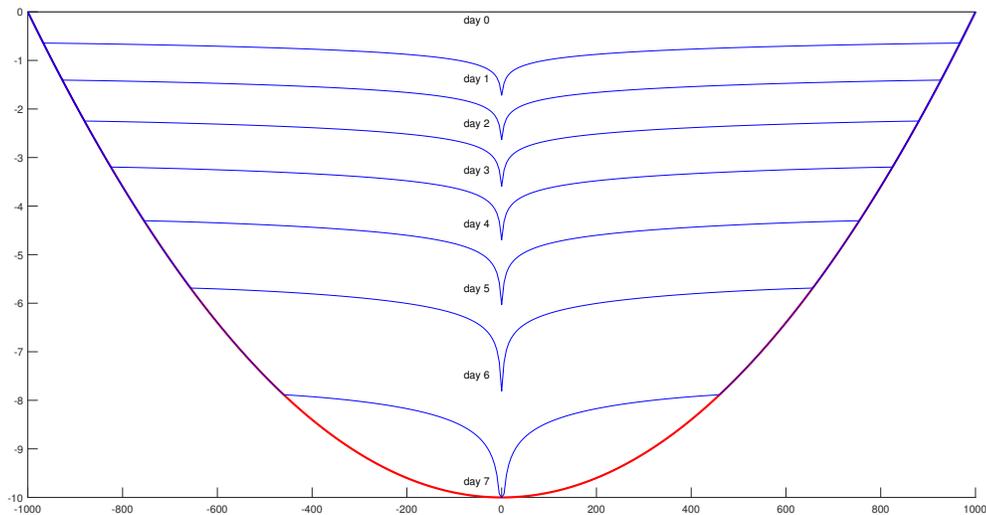


Figure 3: Level of water at days  $0, 1, \dots, 7$  leaking from a pointwise sink at the bottom of a paraboloid phreatic aquifer.

solute value equations and minimizing a quadratic function over the non-negative orthant. A semi-smooth Newton method has been proposed for this problem, where the existence and uniqueness of solutions were studied together with the finite convergence of the method. In [2], the authors conjecture that positive definiteness of  $T$  would be sufficient for finite convergence of the semi-smooth Newton method. However, we showed that this assumption is enough only to avoid cycles of size two in general.

To avoid solving a full linear system of equations at each Newtonian iteration, we proposed Newtonian methods inspired by the classical Jacobi and Gauss-Seidel methods for linear equations, where only a diagonal or triangular linear system is solved at each iteration. The existence and uniqueness of solutions are shown together with global convergence of the methods under stronger variants of the well-known sufficient conditions of convergence for linear systems, namely, diagonal dominance (for the Jacobi iterate) and Sassenfeld's criterion (for the Gauss-Seidel iterate). Numerical experiments were conducted on random problems to attest that the methods are comparable with the standard Newtonian approach, being considerably faster for large-scale and sparse problems. In an applied experiment concerning the discretization of the Boussinesq equation, we show that both methods are robust and reliable even when sufficient conditions for convergence are not met.

For future work, we expect to address the possibility of weakening the sufficient conditions we obtained for the global convergence of the Jacobi-Newton and Gauss-Seidel-Newton iterations. For instance, extensive numerical experiments suggest that Gauss-Seidel-Newton converges globally when  $T$  is a symmetric and positive definite matrix. Another possibility would be to combine Jacobi and Gauss-Seidel iterates in an SOR-style, which may produce interesting theoretical and numerical results. Additionally, instead of considering projection onto the non-negative orthant, we expect to address the analogous equations obtained by projecting onto the second-order cone or the semidefinite cone. The situation is more challenging as the projection matrices in those cases do not have such a simple diagonal structure.

## References

- [1] J. Chen and R. P. Agarwal, “On Newton-type approach for piecewise linear systems,” *Linear Algebra and its Applications*, vol. 433, pp. 1463–1471, Dec. 2010.
- [2] J.-Y. Bello-Cruz, O. P. Ferreira, S. Z. Németh, and L. F. Prudente, “A semi-smooth Newton method for projection equations and linear complementarity problems with respect to the second order cone,” *Linear Algebra Its Appl.*, vol. 513, pp. 160–181, 2017.
- [3] A. Griewank, J.-U. Bernt, M. Radons, and T. Streubel, “Solving piecewise linear systems in abs-normal form,” *Linear Algebra and its Applications*, vol. 471, pp. 500–530, Apr. 2015.
- [4] Z. Sun, L. Wu, and Z. Liu, “A damped semismooth Newton method for the Brugnano–Casulli piecewise linear system,” *Bit Numer Math*, vol. 55, pp. 569–589, June 2015.
- [5] O. L. Mangasarian, “A generalized Newton method for absolute value equations,” *Optim Lett*, vol. 3, pp. 101–108, Jan. 2009.
- [6] O. P. Ferreira and S. Z. Németh, “Projection onto simplicial cones by a semi-smooth Newton method,” *Optim. Lett.*, vol. 9, no. 4, pp. 731–741, 2015.
- [7] J. Barrios, O. P. Ferreira, and S. Z. Németh, “Projection onto simplicial cones by Picard’s method,” *Linear Algebra Appl.*, vol. 480, pp. 27–43, 2015.
- [8] O. L. Mangasarian and R. R. Meyer, “Absolute value equations,” *Linear Algebra and its Applications*, vol. 419, pp. 359–367, Dec. 2006.
- [9] J. G. Barrios, J. Y. Bello Cruz, O. P. Ferreira, and S. Z. Németh, “A semi-smooth Newton method for a special piecewise linear system with application to positively constrained convex quadratic programming,” *J. Comput. Appl. Math.*, vol. 301, pp. 91–100, 2016.
- [10] L. Brugnano and V. Casulli, “Iterative solution of piecewise linear systems and applications to flows in porous media,” *SIAM J. Sci. Comput.*, vol. 31, no. 3, pp. 1858–1873, 2009.
- [11] J. Y. Bello Cruz, O. P. Ferreira, and L. F. Prudente, “On the global convergence of the inexact semi-smooth Newton method for absolute value equation,” *Comput. Optim. Appl.*, vol. 65, no. 1, pp. 93–108, 2016.
- [12] L. Brugnano and V. Casulli, “Iterative solution of piecewise linear systems,” *SIAM Journal on Scientific Computing*, vol. 30, no. 1, pp. 463–472, 2008.
- [13] J. M. Ortega, *Numerical Analysis: A Second Course*. Philadelphia: Society for Industrial and Applied Mathematics, Jan. 1987.
- [14] T. A. Davis, *Direct methods for sparse linear systems*. SIAM, 2006.
- [15] I. S. Duff, A. M. Erisman, and J. K. Reid, *Direct methods for sparse matrices*. Oxford University Press, 2017.
- [16] E. D. Dolan and J. J. Moré, “Benchmarking optimization software with performance profiles,” *Mathematical programming*, vol. 91, no. 2, pp. 201–213, 2002.
- [17] J. Bear and A. Verruijt, “Modeling groundwater flow and pollution,” *EOS*, 1988.