Journal of Computational and Applied Mathematics 236 (2012) 3937–3947

Contents lists available at SciVerse ScienceDirect

Journal of Computational and Applied Mathematics

journal homepage: www.elsevier.com/locate/cam

Iterative solutions of mildly nonlinear systems

Vincenzo Casulli [a,*], Paola Zanolli [b]

[a] *Laboratory of Applied Mathematics, Department of Civil and Environmental Engineering, University of Trento, 38050 Mesiano, Italy*

[b] *Mathematics Department, University of Trento, 38050 Povo, Italy*

ARTICLE INFO

Article history:
Received 4 November 2011
Received in revised form 25 February 2012

Keywords:
Mildly nonlinear systems
Functions of bounded variations
Nested iterations
Free-surface hydrodynamics
Wetting and drying
Confined–unconfined aquifers

ABSTRACT

The correct numerical modelling of free-surface hydrodynamics often requires the solution of diagonally nonlinear systems. In doing this, one may substantially enhance the model accuracy while fulfilling relevant physical constraints. This is the case when a suitable semi-implicit discretization is used, e.g., to solve the one-dimensional or the multi-dimensional shallow water equations; to model axially symmetric flows in compliant arterial systems; to solve the Boussinesq equation in confined–unconfined aquifers; or to solve the mixed form of the Richards equation. In this paper two nested iterative methods for solving a *mildly nonlinear system* of the form $\mathbf{V}(\boldsymbol{\eta}) + T\boldsymbol{\eta} = \mathbf{b}$ are proposed and analysed. It is shown that the inner and the outer iterates are monotone, and converge to the exact solution for a wide class of mildly nonlinear systems of applied interest. A simple, and yet non-trivial test problem derived from the mathematical modelling of flows in porous media is formulated and solved with the proposed methods.

1. Introduction

The problem to be solved is that of finding a solution to the following mildly nonlinear system:

$$\mathbf{V}(\boldsymbol{\eta}) + T\boldsymbol{\eta} = \mathbf{b} \tag{1}$$

in which:

- $\boldsymbol{\eta} \in \mathbb{R}^N$ is unknown.
- $\mathbf{V}$ is a vectorial function $\mathbf{V}(\boldsymbol{\eta}) = (V_i(\eta_i))_{i=1}^N$, where the $V_i(\eta_i)$ are defined for all $\eta_i \in \mathbb{R}$ and can be expressed as

$$V_i(\eta_i) = \int_{-\infty}^{\eta_i} a_i(z)dz. \tag{2}$$

For all $i = 1, 2, \ldots, N$, the following assumptions are made on the function $a_i(\eta)$:

A1: $a_i(\eta)$ is defined for all $\eta \in \mathbb{R}$, and is a non-negative function with bounded variations;

A2: there exist ℓ_i, $u_i \in \mathbb{R}$ such that $a_i(\eta)$ is non-decreasing in $(-\infty, \ell_i]$, and non-increasing in $[u_i, +\infty)$.

- $T \in \mathbb{R}^{N\times N}$ is a symmetric and (at least) positive semidefinite matrix satisfying one of the following properties [1]:

T1: T is a Stieltjes matrix, i.e., a symmetric M-matrix, or

T2: T is irreducible, null$(T) \equiv$ span$(\mathbf{v})$, with $\mathbf{v} > \mathbf{0}$ (componentwise), and $T + D$ is a Stieltjes matrix for all diagonal matrices $D \gneqq O$, with O denoting the null matrix.

[*] Corresponding author.
E-mail address: vincenzo.casulli@unitn.it (V. Casulli).

0377-0427/$ – see front matter © 2012 Elsevier B.V. All rights reserved.
doi:10.1016/j.cam.2012.02.042

- $\mathbf{b} \in \mathbb{R}^N$ is a known vector. When T is **T2**, the following compatibility assumption is required on **b**:

$$0 < \mathbf{v}^\top \mathbf{b} < \mathbf{v}^\top \mathbf{V}^{\text{Max}}, \tag{3}$$

where the entries of $\mathbf{V}^{\text{Max}} \in \mathbb{R}^N$ are given by $V_i^{\text{Max}} = \int_{-\infty}^{+\infty} a_i(z)dz$.

As an example, a particular form for system (1) is obtained when $\mathbf{V}(\boldsymbol{\eta}) = \max(\mathbf{0}, \boldsymbol{\eta})$. In this case, for all $i = 1, 2, \ldots, N$, $V_i(\eta_i)$ can be expressed in the form (2) with $a_i(\eta)$ being a step function given by

$$a_i(\eta) = \begin{cases} 1 & \text{if } \eta \geq 0 \\ 0 & \text{otherwise.} \end{cases} \tag{4}$$

Furthermore, assumptions **A1** and **A2** are satisfied by taking any $\ell_i \in \mathbb{R}$ and $u_i \geq 0$ for all $i = 1, 2, \ldots, N$. The resulting system, called a *piecewise linear system*, can be solved exactly within a finite number of *Newton-type* iterations (see Ref. [2] for details).

When $\mathbf{V}(\boldsymbol{\eta}) = \max[\mathbf{0}, \min(\mathbf{1}, \boldsymbol{\eta})]$, one gets another example of a *piecewise linear system*. Here $V_i(\eta_i)$ can be expressed in the form (2) with $a_i(\eta)$ given by

$$a_i(\eta) = \begin{cases} 1 & \text{if } 0 \leq \eta \leq 1 \\ 0 & \text{otherwise.} \end{cases} \tag{5}$$

Furthermore, assumptions **A1** and **A2** are satisfied on taking $\ell_i \leq 1$ and $u_i \geq 0$ for all $i = 1, 2, \ldots, N$. The resulting system can be solved exactly within a finite number of iterations by *nested* Newton-type algorithms recently proposed in Ref. [3].

In general, a direct application of Newton-type methods to system (1) may fail to converge unless the initial guess is sufficiently 'close' to the unknown solution [4]. Specific nonlinear extensions of the formulation (4) yield mildly nonlinear systems that can be solved by a Newton-type algorithm with global convergence properties. This method has been used to obtain a high-resolution wetting and drying model for free-surface hydrodynamics [5–7]. Additionally, specific nonlinear extensions of the formulation (5) have been investigated in Ref. [8] with a view to providing a converging iterative algorithm for solving Richards' equation in mixed form.

In this paper the *nested* Newton-type algorithms introduced in Refs. [2,3] for piecewise linear systems are further generalized to solve a wider class of mildly nonlinear systems that are, or can be formulated, as (1), and which arise from several non-trivial problems of applied interest. Sufficient conditions for global convergence, and a quadratic convergence rate are investigated. System (1) is first analysed in Section 2. Then, a nested Newton-type algorithm is derived in Section 3. Section 4 provides a convergence analysis. A dual algorithm is outlined in Section 5. Finally, in Section 6, the proposed algorithms are applied to solve the Boussinesq equation in confined–unconfined aquifers.

2. System analysis

Having assumed that the $a_i(\eta)$ are non-negative functions of bounded variations, they are *almost everywhere differentiable*, admit only *discontinuities of the first kind*, and can be expressed as the difference of two non-negative, bounded, and non-decreasing functions (the *Jordan decomposition* [9]), say $p_i(\eta)$ and $q_i(\eta)$, such that $a_i(\eta) = p_i(\eta) - q_i(\eta) \geq 0$ with $0 = q_i(\ell_i) \leq q_i(\eta) \leq p_i(\eta) \leq p_i(u_i)$ $\forall \eta \in \mathbb{R}$. Furthermore, $\mathbf{V}(\boldsymbol{\eta})$ can also be written as

$$\mathbf{V}(\boldsymbol{\eta}) = \mathbf{V}_1(\boldsymbol{\eta}) - \mathbf{V}_2(\boldsymbol{\eta}), \tag{6}$$

where each component of $\mathbf{V}_1(\boldsymbol{\eta})$ and $\mathbf{V}_2(\boldsymbol{\eta})$ is defined by

$$V_{1,i}(\eta_i) = \int_{-\infty}^{\eta_i} p_i(z)dz \quad \text{and} \quad V_{2,i}(\eta_i) = \int_{-\infty}^{\eta_i} q_i(z)dz. \tag{7}$$

Let $\boldsymbol{\ell} = (\ell_i)_{i=1}^N$, $\boldsymbol{u} = (u_i)_{i=1}^N$, and let $P(\boldsymbol{\eta})$ and $Q(\boldsymbol{\eta})$ denote diagonal matrices whose diagonal entries are $p_i(\eta_i)$ and $q_i(\eta_i)$, respectively. From the above settings one has

$$\forall \boldsymbol{\eta} \in \mathbb{R}^N: \quad \mathbf{0} \leq \mathbf{V}(\boldsymbol{\eta}) \leq \mathbf{V}^{\text{Max}}, \qquad O = Q(\boldsymbol{\ell}) \leq Q(\boldsymbol{\eta}) \leq P(\boldsymbol{\eta}) \leq P(\boldsymbol{u}) \tag{8}$$

$$\forall \boldsymbol{\eta} \leq \boldsymbol{\ell}: \quad Q(\boldsymbol{\eta}) = O, \qquad \mathbf{V}_2(\boldsymbol{\eta}) = \mathbf{0} \tag{9}$$

$$\forall \boldsymbol{\eta} \geq \boldsymbol{u}: \quad P(\boldsymbol{\eta}) = P(\boldsymbol{u}), \qquad \mathbf{V}_1(\boldsymbol{\eta}) = \mathbf{V}_1(\boldsymbol{u}) + P(\boldsymbol{u})(\boldsymbol{\eta} - \boldsymbol{u}). \tag{10}$$

Lemma 1. *Let $a_i(\eta)$ satisfy the assumptions* **A1** *and* **A2**, *and let $p_i(\eta)$ and $q_i(\eta)$ be the Jordan decomposition of $a_i(\eta)$, $i = 1, 2, \ldots, N$. Then, $\forall \boldsymbol{x}, \mathbf{y} \in \mathbb{R}^N$ one has*

$$P(\boldsymbol{x})(\boldsymbol{x} - \mathbf{y}) - [\mathbf{V}_1(\boldsymbol{x}) - \mathbf{V}_1(\mathbf{y})] \geq \mathbf{0} \tag{11}$$

$$Q(\boldsymbol{x})(\boldsymbol{x} - \mathbf{y}) - [\mathbf{V}_2(\boldsymbol{x}) - \mathbf{V}_2(\mathbf{y})] \geq \mathbf{0}. \tag{12}$$

Proof. Inequalities (11) and (12) are easily proved, componentwise, by considering separately the cases $x_i < y_i$ and $x_i > y_i$. □

As an example, when $a_i(\eta)$ are given by (4), one has $p_i \equiv a_i$ and $q_i \equiv 0$. Consequently, $V_{1,i}(\eta_i) = V_i(\eta_i) = \max(0, \eta_i)$, $V_{2,i} \equiv 0$, $V_i^{\text{Max}} = \infty$, and $p_i(u_i) = 1$, for all $i = 1, 2, \ldots, N$. Accordingly, the second inequality of (3) is always satisfied for all $\mathbf{b} \in \mathbb{R}^N$.

As a second example, when the $a_i(\eta)$ are given by (5), one has

$$p_i(\eta) = \begin{cases} 1 & \text{if } \eta \geq 0 \\ 0 & \text{otherwise} \end{cases} \qquad q_i(\eta) = \begin{cases} 1 & \text{if } \eta > 1 \\ 0 & \text{otherwise.} \end{cases}$$

Consequently, $V_{1,i}(\eta_i) = \max(0, \eta_i)$, $V_{2,i}(\eta_i) = \max(0, \eta_i - 1)$, and so $V_i(\eta_i) = V_{1,i}(\eta_i) - V_{2,i}(\eta_i) = \max[0, \min(1, \eta_i)]$, and $V_i^{\text{Max}} = p_i(u_i) = 1$, for all $i = 1, 2, \ldots, N$.

Regarding matrix T, this is typically sparse and very large. Moreover, within the prescribed assumptions, one can easily verify the following results (see, e.g., [2]):

- if T is **T1**, then $T + D$ is Stieltjes and $(T + D)^{-1} \geq O$ for all diagonal matrices $D \geq O$;
- if T is **T2**, then $\mathbf{v}^\top T = \mathbf{0}^\top$; moreover, $T + D$ is Stieltjes and $(T + D)^{-1} \geq O$ for all diagonal matrices $D \gneqq O$.

Regarding the compatibility assumption (3), note that (8) implies $0 \leq \mathbf{v}^\top \mathbf{V}(\boldsymbol{\eta}) \leq \mathbf{v}^\top \mathbf{V}^{\text{Max}}$. Furthermore, if T is **T2** and $\boldsymbol{\eta}$ is a solution of system (1), then $\mathbf{v}^\top \mathbf{V}(\boldsymbol{\eta}) = \mathbf{v}^\top \mathbf{b}$. Consequently, $0 \leq \mathbf{v}^\top \mathbf{b} \leq \mathbf{v}^\top \mathbf{V}^{\text{Max}}$ is a requirement for system (1) to have a solution. Specifically, if $\boldsymbol{\eta} \in \mathbb{R}^N$ is a solution of (1) and:

- $\mathbf{v}^\top \mathbf{b} = 0$, then $\mathbf{v}^\top \mathbf{V}(\boldsymbol{\eta}) = 0$ implying $\mathbf{V}(\boldsymbol{\eta}) = \mathbf{0}$. Accordingly, $\boldsymbol{\eta}$ is also a solution of the linear and singular system $T\boldsymbol{\eta} = \mathbf{b}$. In this case $\boldsymbol{\eta}$ can be determined from $\boldsymbol{\eta} = \boldsymbol{x} - \lambda\mathbf{v}$ where $\boldsymbol{x}$ is a solution of $T\boldsymbol{x} = \mathbf{b}$ and λ is a scalar sufficiently large that $\mathbf{V}(\boldsymbol{x} - \lambda\mathbf{v}) = 0$.
- $\mathbf{v}^\top \mathbf{b} = \mathbf{v}^\top \mathbf{V}^{\text{Max}}$, then $\mathbf{v}^\top \mathbf{V}(\boldsymbol{\eta}) = \mathbf{v}^\top \mathbf{V}^{\text{Max}}$ implying $\mathbf{V}(\boldsymbol{\eta}) = \mathbf{V}^{\text{Max}}$. Accordingly, $\boldsymbol{\eta}$ is also a solution of the linear and singular system $T\boldsymbol{\eta} = \mathbf{b} - \mathbf{V}^{\text{Max}}$. In this case $\boldsymbol{\eta}$ can be determined from $\boldsymbol{\eta} = \boldsymbol{x} + \lambda\mathbf{v}$ where $\boldsymbol{x}$ is a solution of $T\boldsymbol{x} = \mathbf{b} - \mathbf{V}^{\text{Max}}$ and λ is a scalar sufficiently large that $\mathbf{V}(\boldsymbol{x} + \lambda\mathbf{v}) = \mathbf{V}^{\text{Max}}$.

The existence of a solution for system (1) in the remaining non-trivial cases where either T is **T1**, or T is **T2** and the compatibility assumption (3) is satisfied, will be established constructively in the following sections. Additionally, it can be shown that if T is **T1**, then system (1) has a unique solution for all $\mathbf{b} \in \mathbb{R}^N$, whereas if T is **T2** and the compatibility assumption (3) is satisfied, then a solution of system (1) exists, but may not be unique (see, e.g., [3]).

3. Nested iterations

Note first that because of (6), system (1) can also be written as

$$\mathbf{V}_1(\boldsymbol{\eta}) - \mathbf{V}_2(\boldsymbol{\eta}) + T\boldsymbol{\eta} = \mathbf{b}. \tag{13}$$

A nested Newton-type method for solving (13) is derived by linearizing separately, and in order, $\mathbf{V}_2(\boldsymbol{\eta})$ and $\mathbf{V}_1(\boldsymbol{\eta})$. Specifically, a sequence of *outer* iterates $\{\boldsymbol{\eta}^n\}$ is obtained from (13) by linearizing $\mathbf{V}_2(\boldsymbol{\eta})$ as follows:

$$\mathbf{V}_1(\boldsymbol{\eta}^n) - [\mathbf{V}_2(\boldsymbol{\eta}^{n-1}) + Q(\boldsymbol{\eta}^{n-1})(\boldsymbol{\eta}^n - \boldsymbol{\eta}^{n-1})] + T\boldsymbol{\eta}^n = \mathbf{b}.$$

Thus, on choosing $\boldsymbol{\eta}^0 \leq \boldsymbol{\ell}$, the outer iterates $\boldsymbol{\eta}^n$ are determined from

$$\mathbf{V}_1(\boldsymbol{\eta}^n) + \left(T - Q^{n-1}\right)\boldsymbol{\eta}^n = \mathbf{d}^{n-1}, \quad n = 1, 2, \ldots, \tag{14}$$

where $\mathbf{d}^{n-1} = \mathbf{b} + \mathbf{V}_2^{n-1} - Q^{n-1}\boldsymbol{\eta}^{n-1}$, with $\mathbf{V}_2^{n-1} = \mathbf{V}_2(\boldsymbol{\eta}^{n-1})$ and $Q^{n-1} = Q(\boldsymbol{\eta}^{n-1})$. The resulting nth (outer) residual is derived from (13) and is given by

$$\mathbf{r}^n = \mathbf{V}_1(\boldsymbol{\eta}^n) - \mathbf{V}_2(\boldsymbol{\eta}^n) + T\boldsymbol{\eta}^n - \mathbf{b}. \tag{15}$$

Next, in order to determine $\boldsymbol{\eta}^n$ for all $n = 1, 2, \ldots$, a sequence of *inner* iterates $\{\boldsymbol{\eta}^{n,m}\}$ is obtained from (14) by linearizing $\mathbf{V}_1(\boldsymbol{\eta})$ as follows:

$$[\mathbf{V}_1(\boldsymbol{\eta}^{n,m-1}) + P(\boldsymbol{\eta}^{n,m-1})(\boldsymbol{\eta}^{n,m} - \boldsymbol{\eta}^{n,m-1})] + \left(T - Q^{n-1}\right)\boldsymbol{\eta}^{n,m} = \mathbf{d}^{n-1}.$$

Thus, on choosing $\boldsymbol{\eta}^{n,0} \geq \boldsymbol{u}$, the inner iterates $\boldsymbol{\eta}^{n,m}$ are determined from the linear system

$$\left(T + P^{n,m-1} - Q^{n-1}\right)\boldsymbol{\eta}^{n,m} = \boldsymbol{f}^{n,m-1}, \quad m = 1, 2, \ldots, \tag{16}$$

where $\boldsymbol{f}^{n,m-1} = P^{n,m-1}\boldsymbol{\eta}^{n,m-1} - \mathbf{V}_1^{n,m-1} + \mathbf{d}^{n-1}$, with $P^{n,m-1} = P(\boldsymbol{\eta}^{n,m-1})$ and $\mathbf{V}_1^{n,m-1} = \mathbf{V}_1(\boldsymbol{\eta}^{n,m-1})$. The resulting (n, m)th inner residual is derived from (14) and is given by

$$\mathbf{r}^{n,m} = \mathbf{V}_1(\boldsymbol{\eta}^{n,m}) + \left(T - Q^{n-1}\right)\boldsymbol{\eta}^{n,m} - \mathbf{d}^{n-1}. \tag{17}$$

Lemma 2. *The inner and the outer residuals, $\mathbf{r}^{n,m}$ and $\mathbf{r}^n$, are given by*

$$\mathbf{r}^{n,m} = P^{n,m-1}(\boldsymbol{\eta}^{n,m-1} - \boldsymbol{\eta}^{n,m}) - [\mathbf{V}_1(\boldsymbol{\eta}^{n,m-1}) - \mathbf{V}_1(\boldsymbol{\eta}^{n,m})] \geq \mathbf{0} \tag{18}$$

$$\mathbf{r}^{n} = -\{Q^{n-1}(\boldsymbol{\eta}^{n-1} - \boldsymbol{\eta}^{n}) - [\mathbf{V}_2(\boldsymbol{\eta}^{n-1}) - \mathbf{V}_2(\boldsymbol{\eta}^{n})]\} \leq \mathbf{0} \tag{19}$$

and are componentwise non-negative and non-positive, respectively.

Proof. From (17) and (16), one has the equality in (18). Likewise, from (15) and (14) the equality in (19) is derived. The inequalities in both (18) and (19) result from Lemma 1. □

The inner and outer iterations are terminated when $\|\mathbf{r}^{n,m-1}\| < \epsilon$, and $\|\mathbf{r}^{n}\| < \epsilon$, respectively, with ϵ being a sufficiently small prefixed tolerance.

The nested iterative scheme (14) and (16) is summarized into Algorithm 1. This algorithm is a nonlinear extension of the nested Newton-type method presented in Ref. [3].

Algorithm 1

```
Input $\mathbf{V}_1, \mathbf{V}_2, P, Q, \boldsymbol{\ell}, \boldsymbol{u}, T, \mathbf{b}$, and $\epsilon$
Set $\boldsymbol{\eta}^0 \leq \boldsymbol{\ell}$
Do $n = 1, 2, \ldots$
    Set $\boldsymbol{\eta}^{n,0} \geq \boldsymbol{u}$
    Do $m = 1, 2, \ldots$
        Solve $\left(T + P^{n,m-1} - Q^{n-1}\right) \boldsymbol{\eta}^{n,m} = \boldsymbol{f}^{n,m-1}$
        If $\|\mathbf{r}^{n,m}\| < \epsilon$, then set $\boldsymbol{\eta}^n = \boldsymbol{\eta}^{n,m}$ and exit
    End Do
    If $\|\mathbf{r}^{n}\| < \epsilon$, then set $\boldsymbol{\eta} = \boldsymbol{\eta}^n$ and exit
End Do
Output $\boldsymbol{\eta}$
```

4. Convergence

In order to show that Algorithm 1 is well defined and, in exact arithmetic, converges to an exact solution of system (1) as $\epsilon \to 0$, some preliminary results are outlined first.

Lemma 3. *For any $n, m \geq 1$, let $\boldsymbol{\eta}^{n,m}$ and $\boldsymbol{\eta}^{n,m+1}$ be two subsequent inner iterates obtained from (16). If $T + P^{n,m} - Q^{n-1}$ is Stieltjes, then*

$$\boldsymbol{\eta}^{n,m+1} \leq \boldsymbol{\eta}^{n,m}. \tag{20}$$

Proof. Consider two subsequent inner iterates from (16):

$$\left(T + P^{n,m} - Q^{n-1}\right) \boldsymbol{\eta}^{n,m+1} - P^{n,m}\boldsymbol{\eta}^{n,m} + \mathbf{V}_1^{n,m} = \mathbf{d}^{n-1}$$

$$\left(T + P^{n,m-1} - Q^{n-1}\right) \boldsymbol{\eta}^{n,m} - P^{n,m-1}\boldsymbol{\eta}^{n,m-1} + \mathbf{V}_1^{n,m-1} = \mathbf{d}^{n-1}.$$

By equating the left-hand sides, after using Lemma 2, one obtains

$$\left(T + P^{n,m} - Q^{n-1}\right)\left(\boldsymbol{\eta}^{n,m+1} - \boldsymbol{\eta}^{n,m}\right) = -\mathbf{r}^{n,m} \leq \mathbf{0}.$$

Thus, since $T + P^{n,m} - Q^{n-1}$ is Stieltjes, one has $\left(T + P^{n,m} - Q^{n-1}\right)^{-1} \geq O$ and, consequently, $\boldsymbol{\eta}^{n,m} \geq \boldsymbol{\eta}^{n,m+1}$. □

Lemma 4. *For any $n, m \geq 1$, let $\boldsymbol{\eta}^n$ be the nth outer iterate obtained from (14), and let $\boldsymbol{\eta}^{n+1,m}$ be the subsequent mth inner iterate satisfying (16). If $T + P^{n+1,m-1} - Q^n$ is Stieltjes, then*

$$\boldsymbol{\eta}^{n+1,m} \geq \boldsymbol{\eta}^{n}. \tag{21}$$

Proof. From (16) one has that $\boldsymbol{\eta}^{n+1,m}$ is a solution of

$$\left(T + P^{n+1,m-1} - Q^{n}\right) \boldsymbol{\eta}^{n+1,m} - P^{n+1,m-1}\boldsymbol{\eta}^{n+1,m-1} + \mathbf{V}_1^{n+1,m-1} = \mathbf{d}^{n}, \tag{22}$$

whereas $\boldsymbol{\eta}^n$ is a solution of (14). By subtracting (14) from (22), after using Lemmas 1 and 2, one has

$$\left(T + P^{n+1,m-1} - Q^{n}\right)\left(\boldsymbol{\eta}^{n+1,m} - \boldsymbol{\eta}^{n}\right) = P^{n+1,m-1}(\boldsymbol{\eta}^{n+1,m-1} - \boldsymbol{\eta}^{n}) - (\mathbf{V}_1^{n+1,m-1} - \mathbf{V}_1^{n}) - \mathbf{r}^n \geq \mathbf{0}.$$

Thus, since $T + P^{n+1,m-1} - Q^n$ is Stieltjes, one has $\left(T + P^{n+1,m-1} - Q^{n}\right)^{-1} \geq O$ and, consequently, $\boldsymbol{\eta}^{n+1,m} \geq \boldsymbol{\eta}^{n}$. □

Lemma 5. *If matrix T is* **T2***, the vector function $\mathbf{V}(\boldsymbol{\eta})$ and the vector $\mathbf{b}$ satisfy the hypothesis (3), and if $\boldsymbol{\eta}^n$ is the nth outer iterate obtained from (14), then*

$$\mathbf{v}^\top \mathbf{r}^n = \mathbf{v}^\top \mathbf{V}(\boldsymbol{\eta}^n) - \mathbf{v}^\top \mathbf{b} \leq 0. \tag{23}$$

Proof. Inequality (23) is an immediate consequence of property **T2** and Lemma 2. □

The following results show that Algorithm 1 is well defined, monotone, and converging, in both cases where T is either **T1**, or **T2**. The two cases are analysed separately.

Theorem 1. *Let T be* **T1***. If both assumptions* **A1** *and* **A2** *are satisfied, then for all n and m, matrix $T + P^{n,m-1} - Q^{n-1}$ is Stieltjes, Algorithm 1 is well defined, and the inner and the outer iterates are monotone. Specifically, for all $n = 1, 2, \ldots$, one has*

$$\boldsymbol{\eta}^{n,m+1} \leq \boldsymbol{\eta}^{n,m}, \quad m = 1, 2, \ldots \tag{24}$$

$$\boldsymbol{\eta}^{n} \leq \boldsymbol{\eta}^{n+1,m}, \quad m = 1, 2, \ldots \tag{25}$$

$$\boldsymbol{\eta}^{n} \leq \boldsymbol{\eta}^{n+1}. \tag{26}$$

Moreover, the outer iterates $\boldsymbol{\eta}^n$ converge to the exact solution of system (1).

Proof. Assumptions **A1** and **A2** are required to formulate Algorithm 1. One proceeds by double induction.

$n = 1$: In this case one has $Q^0 = Q(\boldsymbol{\eta}^0 \leq \boldsymbol{\ell}) = O$ and $P^{1,m-1} \geq O$. Thus, matrices $T + P^{1,m-1} - Q^0$ are Stieltjes for all $m = 1, 2, \ldots$. Consequently, Lemma 3 implies inequality (24). Hence, the sequence $\{\boldsymbol{\eta}^{1,m}\}$ is monotonically decreasing. Accordingly, the sequence $\{\mathbf{V}_1(\boldsymbol{\eta}^{1,m})\}$ is non-negative and monotonically decreasing. To prove boundedness of the first inner iterates recall that $T^{-1} \geq O$ and let $\boldsymbol{\eta}^b = T^{-1}[\mathbf{b} - \mathbf{V}_1(\boldsymbol{\eta}^{1,1})]$, so from (16) one has

$$\boldsymbol{\eta}^{1,m} - \boldsymbol{\eta}^b = T^{-1}[\mathbf{V}_1(\boldsymbol{\eta}^{1,1}) - \mathbf{V}_1(\boldsymbol{\eta}^{1,m}) + \mathbf{r}^{1,m}] \geq \mathbf{0}.$$

Hence, the monotonically decreasing sequence $\{\boldsymbol{\eta}^{1,m}\}$ is bounded from below by $\boldsymbol{\eta}^b$, and is therefore converging to, say, $\bar{\boldsymbol{\eta}}^1$. Consequently, $\{P^{1,m}\}$ and $\{\mathbf{V}_1(\boldsymbol{\eta}^{1,m})\}$ converge to $P(\bar{\boldsymbol{\eta}}^1)$ and $\mathbf{V}_1(\bar{\boldsymbol{\eta}}^1)$, respectively. Thus, (18) yields that $\{\mathbf{r}^{1,m}\}$ converges to $\mathbf{0}$, implying that $\boldsymbol{\eta}^1 = \bar{\boldsymbol{\eta}}^1$ is an exact solution of (14).

$n > 1$: One assumes that the $(k-1)$th outer cycle has been successfully completed and so $\boldsymbol{\eta}^{n-1}$ is a solution of (14), and $Q^{n-1} \leq P^{n-1}$.

$m = 1$: Since T is **T1**, and because $P^{n,0} = P(\boldsymbol{u}) \geq Q^{n-1}$, one has that $T + P^{n,0} - Q^{n-1}$ is Stieltjes. Hence $\boldsymbol{\eta}^{n,1}$ can be uniquely determined from (16).

$m > 1$: One assumes that $T + P^{n,m-2} - Q^{n-1}$ is Stieltjes, so $\boldsymbol{\eta}^{n,m-1}$ can be uniquely determined from (16) and, because of Lemma 4, it satisfies $\boldsymbol{\eta}^{n,m-1} \geq \boldsymbol{\eta}^{n-1}$. Consequently, $P^{n,m-1} \geq P^{n-1} \geq Q^{n-1}$ implying that $T + P^{n,m-1} - Q^{n-1}$ is Stieltjes, and $\boldsymbol{\eta}^{n,m}$ can be uniquely determined from (16).

This shows that $T + P^{n,m-1} - Q^{n-1}$ are Stieltjes for all $m = 1, 2, \ldots$. Hence, Lemma 3 implies inequality (24), and Lemma 4 yields inequality (25). This latter inequality shows that the monotonically decreasing sequence $\{\boldsymbol{\eta}^{n,m}\}$ is bounded from below by $\boldsymbol{\eta}^{n-1}$, and therefore is converging to, say, $\bar{\boldsymbol{\eta}}^n$. Consequently, $\{P^{n,m}\}$ and $\{\mathbf{V}_1(\boldsymbol{\eta}^{n,m})\}$ converge to $P(\bar{\boldsymbol{\eta}}^n)$ and $\mathbf{V}_1(\bar{\boldsymbol{\eta}}^n)$, respectively. Thus, from (18), $\mathbf{r}^{n,m}$ converges to $\mathbf{0}$, implying that $\boldsymbol{\eta}^n = \bar{\boldsymbol{\eta}}^n$ is an exact solution of (14). Finally, inequality (26) is an immediate consequence of (25).

To prove convergence of the outer iterates, one also needs to show that the monotonically increasing sequence $\{\boldsymbol{\eta}^n\}$ is bounded from above. With this purpose, let $\boldsymbol{\eta}^t = T^{-1}\mathbf{b}$, and recall that $T^{-1} \geq O$, $\mathbf{V}(\boldsymbol{\eta}^n) \geq 0$ and $\mathbf{r}^n \leq \mathbf{0}$. Hence from (14) one has

$$\boldsymbol{\eta}^t - \boldsymbol{\eta}^n = T^{-1}[\mathbf{V}(\boldsymbol{\eta}^n) - \mathbf{r}^n] \geq \mathbf{0}.$$

Thus, the monotonically increasing sequence $\{\boldsymbol{\eta}^n\}$ is bounded from above by $\boldsymbol{\eta}^t$. Therefore it converges to, say, $\bar{\boldsymbol{\eta}}$. Accordingly, $\{Q^n\}$ and $\{\mathbf{V}_2(\boldsymbol{\eta}^n)\}$ converge to $Q(\bar{\boldsymbol{\eta}})$ and $\mathbf{V}_2(\bar{\boldsymbol{\eta}})$, respectively. Thus, from (19), $\{\mathbf{r}^n\}$ converges to $\mathbf{0}$, implying that $\boldsymbol{\eta} = \bar{\boldsymbol{\eta}}$ is an exact solution of system (1). □

It is interesting to note that Theorem 1 remains true if the initial guess for the inner cycle is relaxed to $\boldsymbol{\eta}^{n,0} \geq \boldsymbol{\eta}^{n-1}$. Thus, for example, by taking $\boldsymbol{\eta}^{n,0} = \boldsymbol{\eta}^{n-1}$ a faster convergence of each inner cycle may be achieved.

Theorem 2. *Let T be an irreducible matrix satisfying* **T2** *and assume that* $\mathbf{b}$ *fulfils the compatibility inequalities (3). If both* **A1** *and* **A2** *are satisfied, then for all n and m, matrix $T + P^{n,m-1} - Q^{n-1}$ is Stieltjes, and Algorithm 1 is well defined and monotone. Specifically, inequalities (24)–(26) are satisfied for all $n = 1, 2, \ldots$, and the outer iterates $\boldsymbol{\eta}^n$ converge to an exact solution of system (1).*

Proof. Assumptions **A1** and **A2** are required to formulate Algorithm 1. Moreover, $P(\boldsymbol{u}) \geq O$ and $P(\boldsymbol{u}) \neq O$. In fact, $P(\boldsymbol{u}) = O$ yields $\mathbf{V}^{\mathrm{Max}} = \mathbf{0}$ which contradicts the compatibility assumption (3). The proof now proceeds by double induction.

$n = 1$: When $n = 1$ one has $Q^{n-1} = Q^0 = Q(\boldsymbol{\eta}^0 \leq \boldsymbol{\ell}) = O$.

$m = 1$: Since $P^{1,0} = P(\boldsymbol{\eta}^{1,0} \geq \boldsymbol{u}) = P(\boldsymbol{u}) \gneqq O$, one has that $T + P^{1,0} - Q^0$ is Stieltjes. Thus, the first inner iterate is well defined.

$m > 1$: One assumes that $T + P^{1,m-1} - Q^0$ is Stieltjes, so $\boldsymbol{\eta}^{1,m}$ can be uniquely determined from (16). Consequently, $P^{1,m} \geq Q^0 = 0$ and $P^{1,m} \neq 0$. Consider, in fact, the inner residual which, for $n = 1$, is given by

$$\mathbf{r}^{1,m} = \mathbf{V}_1(\boldsymbol{\eta}^{1,m}) + T\boldsymbol{\eta}^{1,m} - \mathbf{b} \geq \mathbf{0}.$$

Now, if $P^{1,m} = 0$, one has $\mathbf{V}_1(\boldsymbol{\eta}^{1,m}) = \mathbf{0}$ and hypothesis **T2** implies $\mathbf{v}^\top \mathbf{r}^{1,m} = -\mathbf{v}^\top \mathbf{b} \geq 0$ which contradicts the compatibility assumption (3). This implies $P^{1,m} \neq 0$; thus $T + P^{1,m} - Q^0$ is Stieltjes. Hence, the inner iterates (16) are well defined for all m. Additionally, Lemma 3 implies inequality (24).

To prove boundedness of the first inner iterates, note that inequality (24) implies that the sequence $\{P(\boldsymbol{\eta}^{1,m})\}$ is monotonically decreasing and bounded from below by 0, hence converging to, say, $\bar{P}^1$. Similarly, the sequence $\{\mathbf{V}_1(\boldsymbol{\eta}^{1,m})\}$ is monotonically decreasing and bounded from below by $\mathbf{0}$, hence converging to, say, $\bar{\mathbf{V}}_1^1$. Furthermore, hypothesis **T2** and Lemma 2 imply that the sequence $\{\mathbf{v}^\top \mathbf{r}^{1,m}\}$ is non-negative and monotonically decreasing, and hence bounded. Thus, $\{\mathbf{r}^{1,m}\}$ converges to, say, $\bar{\mathbf{r}}^1 \geq \mathbf{0}$. Accordingly, the sequence $\{T\boldsymbol{\eta}^{1,m}\}$ is itself bounded and converges to $\mathbf{b} + \bar{\mathbf{r}}^1 - \bar{\mathbf{V}}_1^1$.

Now, since T is irreducible, the monotonically decreasing sequence $\{\boldsymbol{\eta}^{1,m}\}$ is either bounded, or componentwise unbounded. Unbounded growth of $\{\boldsymbol{\eta}^{1,m}\}$ to $-\infty$, however, implies $\bar{\mathbf{V}}_1^1 = \mathbf{0}$ and, accordingly, hypothesis **T2** yields $\mathbf{v}^\top \bar{\mathbf{r}}_1 = -\mathbf{v}^\top \mathbf{b} \geq 0$ which contradicts the compatibility assumption (3). Hence the sequence $\{\boldsymbol{\eta}^{1,m}\}$ must be bounded and therefore converging to, say, $\bar{\boldsymbol{\eta}}^1 \in \mathbb{R}^N$. Thus, (18) yields that $\{\mathbf{r}^{1,m}\}$ converges to $\bar{\mathbf{r}}^1 = \mathbf{0}$, implying that $\boldsymbol{\eta}^1 = \bar{\boldsymbol{\eta}}^1$ is an exact solution of (14).

$n > 1$: One assumes that the $(n-1)$st outer cycle has been successfully completed and so $\boldsymbol{\eta}^{n-1}$ is a solution of (14). Moreover, $Q^{n-1} \leq P(\boldsymbol{u})$ and $Q^{n-1} \neq P(\boldsymbol{u})$. In fact, $Q^{n-1} = P(\boldsymbol{u})$ yields $\mathbf{V}(\boldsymbol{\eta}^{n-1}) = \mathbf{V}^{\mathrm{Max}}$. Consequently Lemma 5 implies $\mathbf{v}^\top \mathbf{r}^{n-1} = \mathbf{v}^\top \mathbf{V}^{\mathrm{Max}} - \mathbf{v}^\top \mathbf{b} \leq 0$ which contradicts the compatibility assumption (3).

$m = 1$: $P^{n,0} = P(\boldsymbol{\eta}^{n,0} \geq \boldsymbol{u}) = P(\boldsymbol{u}) \gneqq Q^{n-1}$ implies that $T + P^{n,0} - Q^{n-1}$ is Stieltjes. Consequently, the first inner iterate is well defined.

$m > 1$: One assumes that $T + P^{n,m-1} - Q^{n-1}$ is Stieltjes, so $\boldsymbol{\eta}^{n,m}$ can be uniquely determined from (16) and, because of Lemma 4, it satisfies $\boldsymbol{\eta}^{n,m} \geq \boldsymbol{\eta}^{n-1}$. Accordingly, $P^{n,m} \geq Q^{n-1}$. Now, if $P^{n,m} \neq Q^{n-1}$, then $T + P^{n,m} - Q^{n-1}$ is Stieltjes and $\boldsymbol{\eta}^{n,m+1}$ can be uniquely determined from (16). Alternatively, $P^{n,m} = Q^{n-1}$ and the corresponding inner residual can be written as

$$\begin{aligned}\mathbf{r}^{n,m} &= \mathbf{V}_1(\boldsymbol{\eta}^{n,m}) + (T - Q^{n-1})\boldsymbol{\eta}^{n,m} - \mathbf{b} - \mathbf{V}_2^{n-1} + Q^{n-1}\boldsymbol{\eta}^{n-1} \\ &= \mathbf{V}(\boldsymbol{\eta}^{n-1}) + T\boldsymbol{\eta}^{n,m} - \mathbf{b} - \left[P^{n,m}\left(\boldsymbol{\eta}^{n,m} - \boldsymbol{\eta}^{n-1}\right) - (\mathbf{V}_1^{n,m} - \mathbf{V}_1^{n-1})\right].\end{aligned}$$

Thus, Lemmas 1 and 5 yield $\mathbf{v}^\top \mathbf{r}^{n,m} \leq 0$, but since $\mathbf{r}^{n,m} \geq 0$, one gets $\mathbf{r}^{n,m} = 0$, implying that $\boldsymbol{\eta}^{n,m}$ is an exact solution of (14). Hence $T + P^{n,m-1} - Q^{n-1}$ are Stieltjes, and the inner iterates (16) are well defined for all m, or until convergence is achieved. Furthermore, Lemma 3 implies inequality (24) and Lemma 4 yields inequality (25). This latter inequality shows that the monotonically decreasing sequence $\{\boldsymbol{\eta}^{n,m}\}$ is bounded from below by $\boldsymbol{\eta}^{n-1}$, and therefore is converging to, say, $\bar{\boldsymbol{\eta}}^n$. Consequently, $\{P^{n,m}\}$ and $\{\mathbf{V}_1(\boldsymbol{\eta}^{n,m})\}$ converge to $P(\bar{\boldsymbol{\eta}}^n)$ and $\mathbf{V}_1(\bar{\boldsymbol{\eta}}^n)$, respectively. Thus, from (18), $\{\mathbf{r}^{n,m}\}$ converges to $\mathbf{0}$, implying that $\boldsymbol{\eta}^n = \bar{\boldsymbol{\eta}}^n$ is an exact solution of (14). Finally, inequality (26) is an immediate consequence of (25).

To prove convergence of the outer iterations, one also needs to show the boundedness of the monotonically increasing sequence $\{\boldsymbol{\eta}^n\}$. To this end, consider that the outer residual satisfies

$$\mathbf{r}^n = \mathbf{V}(\boldsymbol{\eta}^n) + T\boldsymbol{\eta}^n - \mathbf{b} \leq \mathbf{0}.$$

Thus, assumption **T2** implies $\mathbf{v}^\top \mathbf{V}(\boldsymbol{\eta}^n) \leq \mathbf{v}^\top \mathbf{b}$. Hence the non-negative and non-decreasing sequence $\{\mathbf{V}(\boldsymbol{\eta}^n)\}$ is bounded and converging to, say, $\bar{\mathbf{V}}$. Additionally, $\mathbf{v}^\top \mathbf{r}^n = \mathbf{v}^\top \mathbf{V}(\boldsymbol{\eta}^n) - \mathbf{v}^\top \mathbf{b}$ is non-positive and monotonically increasing, and hence bounded. This implies that the sequence $\{\mathbf{r}^n\}$ converges to, say, $\bar{\mathbf{r}}$. Consequently, the sequence $\{T\boldsymbol{\eta}^n\}$ is itself bounded, and converges to $\mathbf{b} + \bar{\mathbf{r}} - \bar{\mathbf{V}}$.

Now, since T is irreducible, the monotonically increasing sequence $\{\boldsymbol{\eta}^n\}$ is either bounded, or componentwise unbounded. Unbounded growth of $\{\boldsymbol{\eta}^n\}$ to $+\infty$ componentwise, however, yields $\bar{\mathbf{V}} = \mathbf{V}^{\mathrm{Max}}$ and, accordingly, hypothesis **T2** implies $\mathbf{v}^\top \bar{\mathbf{r}} = \mathbf{v}^\top[\mathbf{V}^{\mathrm{Max}} - \mathbf{b}] \leq 0$ which contradicts the compatibility assumption (3). Hence, the sequence $\{\boldsymbol{\eta}^n\}$ must be bounded, therefore converging to, say, $\bar{\boldsymbol{\eta}} \in \mathbb{R}^N$. Consequently, $\{Q^n\}$ and $\{\mathbf{V}_2(\boldsymbol{\eta}^n)\}$ converge to $\bar{Q}$ and $\mathbf{V}_2(\bar{\boldsymbol{\eta}})$, respectively. Thus, (19) yields that $\{\mathbf{r}^n\}$ converges to $\mathbf{0}$, implying that $\boldsymbol{\eta} = \bar{\boldsymbol{\eta}}$ is the exact solution of system (1). □

Note that if $P(\boldsymbol{\eta}^{n-1}) \neq Q(\boldsymbol{\eta}^{n-1})$, then the initial guess for the inner cycle can be relaxed to $\boldsymbol{\eta}^{n,0} \geq \boldsymbol{\eta}^{n-1}$. In this case, in fact, Theorem 2 remains true and faster convergence of the nth inner cycle may be achieved by taking, e.g., $\boldsymbol{\eta}^{n,0} = \boldsymbol{\eta}^{n-1}$.

Remark 1. If $\boldsymbol{\eta}^1 \leq \boldsymbol{\ell}$, one has $\mathbf{V}_2(\boldsymbol{\eta}^1) = \mathbf{0}$ and, consequently, $\boldsymbol{\eta} = \boldsymbol{\eta}^1$ is a solution of system (13). In this case convergence of Algorithm 1 is achieved in just one outer iteration.

Remark 2. If $\boldsymbol{\eta}^{n,1} \geq \boldsymbol{u}$ for some $n \geq 1$, one has $\mathbf{V}_1(\boldsymbol{\eta}^{n,1}) = \mathbf{V}_1(\boldsymbol{u}) + P(\boldsymbol{u})(\boldsymbol{\eta}^{n,1} - \boldsymbol{u})$ and, consequently, $\boldsymbol{\eta}^{n,1}$ is a solution of system (14). Thus $\boldsymbol{\eta}^n = \boldsymbol{\eta}^{n,1}$ and convergence of the inner iterations is achieved in one step.

Remark 3. Of course, if $\boldsymbol{u} \leq \boldsymbol{\eta}^{1,1} \leq \boldsymbol{\ell}$, then the previous two remarks can be combined to conclude that $\boldsymbol{\eta} = \boldsymbol{\eta}^{1,1}$ is a solution of system (1) which is obtained with Algorithm 1 in just one inner, and one outer iteration.

In many problems of applied interest, matrix T is sparse and very large. Moreover, since T is symmetric and (at least) positive semidefinite, one has that $T + P^{n,m-1} - Q^{n-1}$ are positive definite for all n and m. Hence, in practice, each linear system in the inner iterates (16) can be efficiently solved by a preconditioned conjugate gradient method (see, e.g., [10]).

5. A dual algorithm

System (1) can also be solved by means of a *dual* algorithm. With this purpose, recall that (1) can also be written as

$$\mathbf{V}_1(\boldsymbol{\eta}) - \mathbf{V}_2(\boldsymbol{\eta}) + T\boldsymbol{\eta} = \mathbf{b}. \tag{27}$$

A dual algorithm for solving system (27) is derived by linearizing separately, and in a reverted order, $\mathbf{V}_1(\boldsymbol{\eta})$ and $\mathbf{V}_2(\boldsymbol{\eta})$. Specifically, a sequence of *outer* iterates $\{\tilde{\boldsymbol{\eta}}^n\}$ is obtained from (27) by linearizing $\mathbf{V}_1(\boldsymbol{\eta})$ as follows:

$$[\mathbf{V}_1(\tilde{\boldsymbol{\eta}}^{n-1}) + P(\tilde{\boldsymbol{\eta}}^{n-1})(\tilde{\boldsymbol{\eta}}^n - \tilde{\boldsymbol{\eta}}^{n-1})] - \mathbf{V}_2(\tilde{\boldsymbol{\eta}}^n) + T\tilde{\boldsymbol{\eta}}^n = \mathbf{b}.$$

Thus, on choosing $\tilde{\boldsymbol{\eta}}^0 \geq \boldsymbol{u}$, the outer iterates $\tilde{\boldsymbol{\eta}}^n$ are determined from

$$\left(T + \tilde{P}^{n-1}\right)\tilde{\boldsymbol{\eta}}^n - \mathbf{V}_2(\tilde{\boldsymbol{\eta}}^n) = \tilde{\mathbf{d}}^{n-1}, \quad n = 1, 2, \ldots, \tag{28}$$

where $\tilde{\mathbf{d}}^{n-1} = \mathbf{b} - \tilde{\mathbf{V}}_1^{n-1} + \tilde{P}^{n-1}\tilde{\boldsymbol{\eta}}^{n-1}$, with $\tilde{\mathbf{V}}_1^{n-1} = \mathbf{V}_1(\tilde{\boldsymbol{\eta}}^{n-1})$ and $\tilde{P}^{n-1} = P(\tilde{\boldsymbol{\eta}}^{n-1})$. The resulting nth (outer) residual is

$$\tilde{\mathbf{r}}^n = \mathbf{V}_1(\tilde{\boldsymbol{\eta}}^n) - \mathbf{V}_2(\tilde{\boldsymbol{\eta}}^n) + T\tilde{\boldsymbol{\eta}}^n - \mathbf{b}. \tag{29}$$

Next, in order to determine $\tilde{\boldsymbol{\eta}}^n$ for all $n = 1, 2, \ldots,$ a sequence of *inner* iterates $\{\tilde{\boldsymbol{\eta}}^{n,m}\}$ is obtained from system (28) by linearizing $\mathbf{V}_2(\tilde{\boldsymbol{\eta}})$ as follows:

$$\left(T + \tilde{P}^{n-1}\right)\tilde{\boldsymbol{\eta}}^{n,m} - [\mathbf{V}_2(\tilde{\boldsymbol{\eta}}^{n,m-1}) + Q(\tilde{\boldsymbol{\eta}}^{n,m-1})(\tilde{\boldsymbol{\eta}}^{n,m} - \tilde{\boldsymbol{\eta}}^{n,m-1})] = \tilde{\mathbf{d}}^{n-1}.$$

Thus, on choosing $\tilde{\boldsymbol{\eta}}^{n,0} \leq \boldsymbol{\ell}$, the inner iterates $\tilde{\boldsymbol{\eta}}^{n,m}$ are determined from the linear system

$$\left(T + \tilde{P}^{n-1} - \tilde{Q}^{n,m-1}\right)\tilde{\boldsymbol{\eta}}^{n,m} = \tilde{\boldsymbol{f}}^{n,m-1}, \quad m = 1, 2, \ldots, \tag{30}$$

where $\tilde{\boldsymbol{f}}^{n,m-1} = \tilde{\mathbf{V}}_2^{n,m-1} - \tilde{Q}^{n,m-1}\tilde{\boldsymbol{\eta}}^{n,m-1} + \tilde{\mathbf{d}}^{n-1}$, with $\tilde{Q}^{n,m-1} = Q(\tilde{\boldsymbol{\eta}}^{n,m-1})$ and $\tilde{\mathbf{V}}_2^{n,m-1} = \mathbf{V}_2(\tilde{\boldsymbol{\eta}}^{n,m-1})$. The resulting (n, m)th inner residual is

$$\tilde{\mathbf{r}}^{n,m} = -\mathbf{V}_2(\tilde{\boldsymbol{\eta}}^{n,m}) + \left(T + \tilde{P}^{n-1}\right)\tilde{\boldsymbol{\eta}}^{n,m} - \tilde{\mathbf{d}}^{n-1}. \tag{31}$$

The inner and outer iterations are terminated when $\|\tilde{\mathbf{r}}^{n,m-1}\| < \epsilon$, and $\|\tilde{\mathbf{r}}^n\| < \epsilon$, respectively.

The nested iterative scheme (28) and (30) is summarized into Algorithm 2. This algorithm is a nonlinear extension of the dual method presented in Ref. [3].

Algorithm 2

```
Input $\mathbf{V}_1, \mathbf{V}_2, P, Q, \boldsymbol{\ell}, \boldsymbol{u}, T, \mathbf{b}$, and $\epsilon$
Set $\tilde{\boldsymbol{\eta}}^0 \geq \boldsymbol{u}$
Do $n = 1, 2, \ldots$
    Set $\tilde{\boldsymbol{\eta}}^{n,0} \leq \boldsymbol{\ell}$
    Do $m = 1, 2, \ldots$
        Solve $\left(T + \tilde{P}^{n-1} - \tilde{Q}^{n,m-1}\right)\tilde{\boldsymbol{\eta}}^{n,m} = \tilde{\boldsymbol{f}}^{n,m-1}$
        If $\|\tilde{\mathbf{r}}^{n,m}\| < \epsilon$, then set $\tilde{\boldsymbol{\eta}}^n = \tilde{\boldsymbol{\eta}}^{n,m}$ and exit
    End Do
    If $\|\tilde{\mathbf{r}}^n\| < \epsilon$, then set $\boldsymbol{\eta} = \tilde{\boldsymbol{\eta}}^n$ and exit
End Do
Output $\boldsymbol{\eta}$
```

Convergence of Algorithm 2 can be established using arguments similar to those given for Algorithm 1. For completeness, the main results are listed next.

Theorem 3. *Let T be* **T1**. *If both assumptions* **A1** *and* **A2** *are satisfied, then for all n and m, matrix $T + \tilde{P}^{n-1} - \tilde{Q}^{n,m-1}$ is Stieltjes, Algorithm 2 is well defined, and the inner and the outer iterates are monotone. Specifically, for all $n = 1, 2, \ldots,$ one has*

$$\tilde{\boldsymbol{\eta}}^{n,m+1} \geq \tilde{\boldsymbol{\eta}}^{n,m}, \quad m = 1, 2, \ldots \tag{32}$$

$$\tilde{\boldsymbol{\eta}}^{n} \geq \tilde{\boldsymbol{\eta}}^{n+1,m}, \quad m = 1, 2, \ldots \tag{33}$$

$$\tilde{\boldsymbol{\eta}}^{n} \geq \tilde{\boldsymbol{\eta}}^{n+1}. \tag{34}$$

Moreover, the outer iterates $\tilde{\boldsymbol{\eta}}^n$ converge to the exact solution of system (1).

Theorem 4. *Let T be an irreducible matrix satisfying* **T2** *and assume that* **b** *fulfils the compatibility inequalities (3). If both* **A1** *and* **A2** *are satisfied, then for all n and m, matrix $T + \tilde{P}^{n-1} - \tilde{Q}^{n,m-1}$ is Stieltjes, and Algorithm 2 is well defined and monotone. Specifically, for all $n = 1, 2, \ldots,$ inequalities (32)–(34) are satisfied and the outer iterates $\tilde{\boldsymbol{\eta}}^n$ converge to an exact solution of system (1).*

The proofs of Theorems 3 and 4 are omitted because they follow in a completely analogous fashion to those of Theorems 1 and 2, respectively.

When either T is **T1**, or T is **T2** and $P(\tilde{\boldsymbol{\eta}}^{n-1}) \neq Q(\tilde{\boldsymbol{\eta}}^{n-1})$, then the initial guess for the inner cycle can be relaxed to $\tilde{\boldsymbol{\eta}}^{n,0} \geq \tilde{\boldsymbol{\eta}}^{n-1}$. In this case faster convergence of the nth inner cycle may be achieved by taking $\tilde{\boldsymbol{\eta}}^{n,0} = \tilde{\boldsymbol{\eta}}^{n-1}$.

Recall that in practical applications, matrix T is often sparse and very large. However, since T is symmetric and (at least) positive semidefinite, matrices $T + \tilde{P}^{n-1} - \tilde{Q}^{n,m-1}$ are positive definite for all n and m. Hence, each linear system in the inner iterates (30) can be efficiently solved by a preconditioned conjugate gradient method (see, e.g., [11]).

Remark 4. If $\tilde{\boldsymbol{\eta}}^1 \geq \boldsymbol{u}$, one has $\mathbf{V}_1(\tilde{\boldsymbol{\eta}}^1) = \mathbf{V}_1(\boldsymbol{u}) + P(\boldsymbol{u})(\tilde{\boldsymbol{\eta}}^1 - \boldsymbol{u})$ and, consequently, $\boldsymbol{\eta} = \tilde{\boldsymbol{\eta}}^1$ is a solution of system (27). In this case convergence of Algorithm 1 is achieved in just one outer iteration.

Remark 5. If $\tilde{\boldsymbol{\eta}}^{n,1} \leq \boldsymbol{\ell}$ for some $n \geq 1$, one has $\mathbf{V}_2(\tilde{\boldsymbol{\eta}}^{n,1}) = \mathbf{0}$ and, consequently, $\tilde{\boldsymbol{\eta}}^{n,1}$ is a solution of system (28). Thus $\tilde{\boldsymbol{\eta}}^n = \tilde{\boldsymbol{\eta}}^{n,1}$ and convergence of the inner iterations is achieved in one step.

Remark 6. Of course, if $\boldsymbol{u} \leq \tilde{\boldsymbol{\eta}}^{1,1} \leq \boldsymbol{\ell}$, the previous two remarks can be combined to conclude that $\boldsymbol{\eta} = \tilde{\boldsymbol{\eta}}^{1,1}$ is a solution of system (1) which is obtained with Algorithm 2 in just one inner, and one outer iteration.

6. Confined–unconfined flows in porous media

Consider the mathematical modelling of a two-dimensional water flow within a horizontal aquifer $\Omega \subset \mathbb{R}^2$. Let the aquifer be delimited below by an impervious bottom, and above by a ceiling, described by the surfaces $z = -h(x, y)$ and $z = c(x, y)$, respectively [3].

Let $a(x, y, z)$ and $\kappa(x, y, z)$ denote the local aquifer *porosity* and *hydraulic conductivity*, respectively. Assume that a and κ are prescribed for all $(x, y) \in \Omega$ and for all $z \in [-h(x, y), c(x, y)]$. For notational convenience, the porosity and the hydraulic conductivity are prolonged as $a(x, y, z) = \kappa(x, y, z) = 0$ for all $z \notin [-h(x, y), c(x, y)]$.

The governing differential equation that applies to both pressurized and phreatic flows is taken to be [12]

$$\frac{\partial}{\partial t}\left[\int_{-h}^{\eta} a\, dz\right] = \frac{\partial}{\partial x}\left[\mathcal{K}\frac{\partial \eta}{\partial x}\right] + \frac{\partial}{\partial y}\left[\mathcal{K}\frac{\partial \eta}{\partial y}\right] + \phi, \tag{35}$$

where: t is the time; $\phi(x, y, t)$ is a source or sink; $\mathcal{K}(x, y, \eta) = \int_{-h}^{\eta} \kappa(x, y, z)dz$; and $\eta(x, y, t)$ is the unknown representing the *piezometric head*.

When and where $\eta(x, y, t) \geq c(x, y)$ the aquifer is confined, and the flow is said to be *pressurized*. If $-h(x, y) < \eta(x, y, t) < c(x, y)$, the surface $z = \eta(x, y, t)$ also represents the *phreatic surface*. Finally, if at any time $\eta(x, y, t) \leq -h(x, y)$ then (x, y) is a *dry* point. Accordingly, the time dependent domain for Eq. (35) is

$$\Omega(t) = \{(x, y) \in \Omega : h(x, y) + \eta(x, y, t) > 0\}.$$

To solve Eq. (35) numerically, Ω is covered by a uniform Cartesian grid with size Δx and Δy. Let $\Omega_{i,j} = [x_{i-\frac{1}{2}}, x_{i+\frac{1}{2}}] \times [y_{j-\frac{1}{2}}, y_{j+\frac{1}{2}}] \cap \Omega$ be a typical control volume and let Δt be the time step. Then, at each time level $t_k = k\Delta t$, a consistent finite volume discretization of Eq. (35) is taken to be

$$\begin{aligned} V_{i,j}(\eta_{i,j}^{k+1}) = V_{i,j}(\eta_{i,j}^{k}) + \Delta t \Bigg[& \frac{\mathcal{D}_{i+\frac{1}{2},j}^{k}\left(\eta_{i+1,j}^{k+1} - \eta_{i,j}^{k+1}\right) - \mathcal{D}_{i-\frac{1}{2},j}^{k}\left(\eta_{i,j}^{k+1} - \eta_{i-1,j}^{k+1}\right)}{\Delta x} \\ & + \frac{\mathcal{D}_{i,j+\frac{1}{2}}^{k}\left(\eta_{i,j+1}^{k+1} - \eta_{i,j}^{k+1}\right) - \mathcal{D}_{i,j-\frac{1}{2}}^{k}\left(\eta_{i,j}^{k+1} - \eta_{i,j-1}^{k+1}\right)}{\Delta y} + \phi_{i,j}^{k}\Bigg], \end{aligned} \tag{36}$$

where

$$V_{i,j}(\eta_{i,j}^{k}) = \int_{-\infty}^{\eta_{i,j}^{k}} a_{i,j}(z)dz, \quad \text{with } a_{i,j}(z) = \int_{\Omega_{i,j}} a(x, y, z)\, dx\, dy \tag{37}$$

and, additionally,

$$\mathcal{D}_{i\pm\frac{1}{2},j}^{k} = \int_{y_{j-\frac{1}{2}}}^{y_{j+\frac{1}{2}}} \mathcal{K}\left(x_{i\pm\frac{1}{2}}, y, \eta_{i\pm\frac{1}{2},j}^{k}\right) dy, \qquad \mathcal{D}_{i,j\pm\frac{1}{2}}^{k} = \int_{x_{i-\frac{1}{2}}}^{x_{i+\frac{1}{2}}} \mathcal{K}\left(x, y_{j\pm\frac{1}{2}}, \eta_{i,j\pm\frac{1}{2}}^{k}\right) dx$$

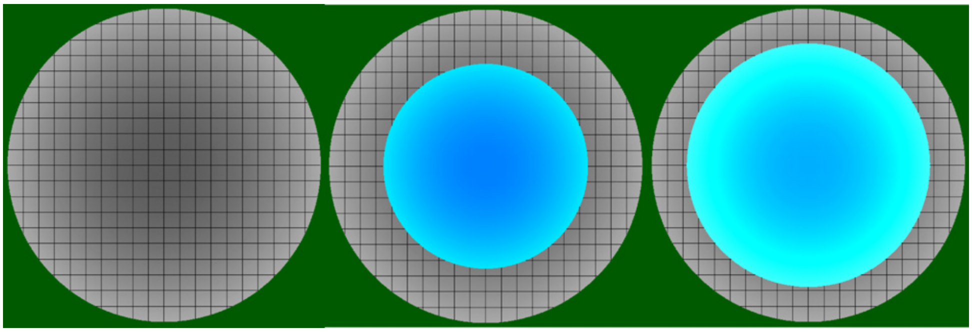

Fig. 1. Pressurized–phreatic region at $t_0 = 0$ (left), at one day (middle), and at two days (right).

with the piezometric heads $\eta^k_{i\pm\frac{1}{2},j}$ and $\eta^k_{i,j\pm\frac{1}{2}}$ between control volumes being defined as averages from the nearest grid values; and, finally,

$$\phi^k_{i,j} = \int_{\Omega_{i,j}} \phi(x, y, t_k)\, dx\, dy.$$

Those control volumes where $\mathcal{D}^k_{i\pm\frac{1}{2},j} = 0$ and $\mathcal{D}^k_{i,j\pm\frac{1}{2}} = 0$ do not contribute to the system that is being formulated. The set of Eqs. (36), where at least one of $\mathcal{D}^k_{i\pm\frac{1}{2},j}$ and $\mathcal{D}^k_{i,j\pm\frac{1}{2}}$ is strictly positive, can be assembled into a *mildly nonlinear* system with unknowns $\eta^{k+1}_{i,j}$. This system (which has to be solved *at every time step*) can be recognized as being in the form (1). The resulting matrix T is sparse, symmetric, and at least positive semidefinite.

Without loss of generality, it will also be assumed that matrix T is irreducible. This may not be the case when, at any time, two or more subdomains are not connected. In such a circumstance the above method applies separately to each such subdomain where the corresponding matrix T is irreducible.

For testing purposes, given two positive constants $h_0 = 10$ m, and $L = 1000$ m, let $\Omega = \{(x, y) : x^2 + y^2 < L^2\}$, and let the aquifer's bottom and ceiling be described by a paraboloid of revolution given by

$$c(x, y) = h(x, y) = h_0 \left(1 - \frac{x^2 + y^2}{L^2}\right).$$

It is also assumed that the aquifer has a constant porosity $a(x, y, z) = 0.3$ and hydraulic conductivity $\kappa(x, y, z) = 1$ m/s. Thus, on setting

$$p(x, y, z) = \begin{cases} 0.3 & \text{if } z \geq -h(x, y) \\ 0 & \text{otherwise} \end{cases} \qquad q(x, y, z) = \begin{cases} 0.3 & \text{if } z > c(x, y) \\ 0 & \text{otherwise} \end{cases}$$

one has $a_{i,j}(z) = p_{i,j}(z) - q_{i,j}(z)$, with $p_{i,j}$ and $q_{i,j}$ being the Jordan decomposition of $a_{i,j}$. Specifically,

$$p_{i,j}(z) = \int_{\Omega_{i,j}} p(x, y, z)\, dx\, dy \quad \text{and} \quad q_{i,j}(z) = \int_{\Omega_{i,j}} q(x, y, z)\, dx\, dy.$$

Finally, in order to specify the initial guess for the outer and for the inner iterations, $\ell_{i,j}$ and $u_{i,j}$ are taken to be

$$\ell_{i,j} = \min_{\Omega_{i,j}} c(x, y) \quad \text{and} \quad u_{i,j} = -\min_{\Omega_{i,j}} h(x, y).$$

As the initial condition, on specifying $\eta(x, y, 0) = h_0$, at the initial time $t_0 = 0$ the aquifer is assumed to be everywhere pressurized. The flow is then driven by a sink, located at the centre of the aquifer that pumps water out at a constant rate $q = 10\ \text{m}^3/\text{s}$.

Then, a numerical simulation is carried out for 10 days by using a rather coarse grid with $\Delta x = \Delta y = 100$ m, and a very large time step $\Delta t = 1$ day. In this test problem, since $\mathcal{D}^k_{i\pm\frac{1}{2},j} = 0$ and $\mathcal{D}^k_{i,j\pm\frac{1}{2}} = 0$ along the outer edges of the computational grid, matrix T is **T2** at all times.

The resulting flow is partially phreatic and partially pressurized until time t_5. The phreatic region is represented by an expanding inner circle in Fig. 1, whereas the pressurized region is represented by the outer gridded ring. From time t_6, as pumping continues, the phreatic region shrinks, leaving space to an outer dry ring of increasing size (see Fig. 2). As shown in Figs. 1 and 2, despite the use of a relatively coarse mesh, the pressurized, the phreatic and the dry regions are well resolved. This is achieved by the present nonlinear formulation (37) that allows an extremely accurate representation of the flow region.

The total water volume at each time level t_k is given by $V^k = \sum_{i,j} V_{i,j}(\eta^k_{i,j})$, and is decreasing at constant rate $q = 10\ \text{m}^3/\text{s}$. For the present setting, since the initial water volume is $V^0 = 9,424,778\ \text{m}^3$, one has $V^k = V^0 - k\Delta t q$. Thus, any attempt

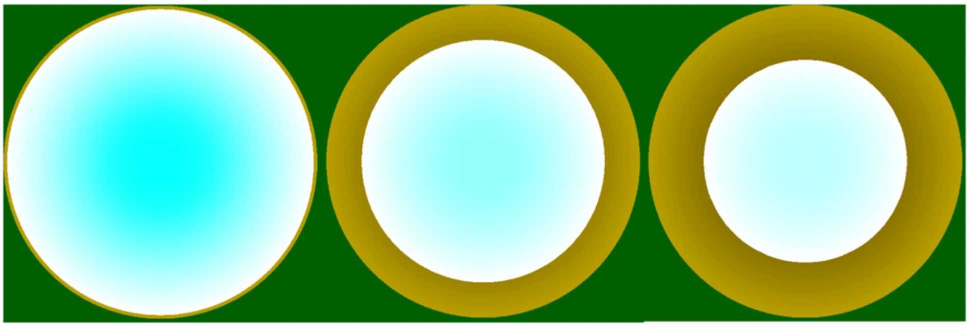

Fig. 2. Wet–dry region at six days (left), at nine days (middle), and at ten days (right).

Table 1
Algorithm 1—problem size, inner, and outer iterations with $\epsilon = 10^{-10}$.

Simulation time	1 *d*	2 *d*	3 *d*	4 *d*	5 *d*	6 *d*	7 *d*	8 *d*	9 *d*	10 *d*
Active control volumes	344	344	344	344	344	332	316	268	216	164
Total inner iterations	5	5	4	4	3	3	4	5	5	5
Outer iterations	5	5	4	4	3	1	1	1	1	1

Table 2
Algorithm 2—problem size, inner, and outer iterations with $\epsilon = 10^{-10}$.

Simulation time	1 *d*	2 *d*	3 *d*	4 *d*	5 *d*	6 *d*	7 *d*	8 *d*	9 *d*	10 *d*
Active control volumes	344	344	344	344	344	332	316	268	216	164
Total inner iterations	5	5	4	4	3	3	4	5	5	5
Outer iterations	1	1	1	1	1	3	4	5	5	5

to extend the simulation beyond t_{10} would produce a *physically unrealistic* negative volume $V^{11} = -79,222$ m^3. On the other hand, when $k = 10$ the compatibility assumption (3) is no longer satisfied and neither Algorithm 1 nor Algorithm 2 will converge. Indeed, when $k = 10$ system (36) does not even have a solution. This demonstrates that the finite volume formulation (36) does not allow for artificial and unphysical overdrainage.

Table 1 shows, for each time level, the size of the resulting system being solved with Algorithm 1, the required number of outer iterations and the total number of inner iterations. The system size remains unchanged during the first five time levels because the aquifer wet domain remains the same. Then, during the subsequent five time levels the system size decreases because dry control volumes do not contribute to the system size. Moreover, it is to be noted that the number of both inner and outer iterations turns out to be remarkably small. Specifically, during the first five time levels each outer iteration requires only one inner iteration for convergence. Then, starting from time level t_6, convergence is achieved with within one outer iteration.

Table 2 shows the system size, the number of outer iterations and the number of total inner iterations required by Algorithm 2. Of course, the mildly nonlinear system being solved at each time step is the same as the one solved by Algorithm 1. Accordingly, the resulting pressurized, phreatic, and dry regions are shown in Figs. 1 and 2. Also, the system size indicated in the first row of Table 2 is identical to the one given in Table 1. The main difference is given by the number of iterations required for convergence. Specifically, the dual algorithm converges in only one outer iteration during the first five time levels. Then, starting from time level t_6, each outer iteration requires only one inner iteration.

7. Conclusions

The iterative solutions of large and sparse mildly nonlinear systems have been considered. After appropriate splitting of the diagonal nonlinear terms, the nonlinear contributions are linearized in sequence to derive a nested iterative method which is shown to be well defined and converging. The splitting is based on a simple Jordan decomposition of functions with bounded variations. A dual algorithm with analogous properties is obtained by reversing the linearization order. Both methods may simplify to the classical Newton method with quadratic convergence rate in several cases of practical interest. Unlike most Newton-type methods, whose convergence is often problematic, the proposed algorithms will always converge when the initial guess is chosen as suggested. These methods apply to a wide class of mildly nonlinear systems that arise from the numerical modelling of free-surface hydrodynamics. As an example, the above algorithms have been applied to simulate porous flows in a confined–unconfined aquifer.

Acknowledgements

It was a great privilege and an immense pleasure to have shared 40 years of deep and sincere friendship with Prof. Donato Trigiante and with his wife Valeria. Donato and Valeria will continue to live in our memory.

References

[1] R.A. Horn, C.R. Johnson, Topics in Matrix Analysis, Cambridge University Press, New York, 1991.
[2] L. Brugnano, V. Casulli, Iterative solution of piecewise linear systems, SIAM J. Sci. Comput. 30 (2008) 463–472.
[3] L. Brugnano, V. Casulli, Iterative solution of piecewise linear systems and applications to porous media, SIAM J. Sci. Comput. 31 (2009) 1858–1873.
[4] J.M. Ortega, W.C. Rheinboldt, Iterative Solution of Nonlinear Equations in Several Variables, Academic Press, New York, NY, 1970.
[5] V. Casulli, A high resolution wetting and drying algorithm for free-surface hydrodynamics, Internat. J. Numer. Methods Fluids 60 (2009) 391–408.
[6] V. Casulli, M. Dumbser, E.F. Toro, Semi-implicit numerical modeling of axially symmetric flows in compliant arterial systems, Int. J. Numer. Methods Biomed. Eng. 28 (2012) 257–272.
[7] V. Casulli, G.S. Stelling, Semi-implicit subgrid modelling of three-dimensional free-surface flows, Internat. J. Numer. Methods Fluids 67 (2011) 441–449.
[8] V. Casulli, P. Zanolli, A nested Newton-type algorithm for finite volume methods solving Richards' equation in mixed form, SIAM J. Sci. Comput. 32 (2010) 2255–2273.
[9] V.V. Chistyakov, On mapping of bounded variation, J. Dyn. Control Syst. 3 (1997) 261–289.
[10] G.H. Golub, C.F. van Loan, Matrix Computations, third ed., The Johns Hopkins University Press, Baltimore, MD, 1996.
[11] Y. Saad, Iterative Methods for Sparse Linear Systems, second ed., SIAM, Philadelphia, 2003.
[12] J. Bear, A. Verruijt, Modeling Groundwater Flow and Pollution, D. Reidel, Dordrecht, Holland, 1987.