CONVERGENCE OF THE MODIFIED PICARD ITERATION FOR THE RICHARDS EQUATION

JUAN B, ANNA MAZZUCATO

Abstract. As part of my thesis work, we investigated convergence of the modified Picard iteration, in order to further our understanding of the problem, and the effect of the nonlinearities on the convergence. In this note, we record a result on the convergence of the Modified Picard iteration, whose proof is adapted from that of the weak convergence of the L-scheme to the unique weak solution of the Cauchy problem.

1. Problem Formulation

Flow of a fluid through porous media is typically considered by applying Darcy’s law, which is essentially a linear relation between the pressure differential in the medium and the flow through it. In many cases, this linear relation is overly simplistic, or the material in question doesn’t satisfy the physical assumptions required for the modeling of flow velocity via Darcy’s Law to be accurate.

One more complex model that is considered in hydrology is Richards’ equation, which allows the computation of the pressure head in poroelastic materials that aren’t saturated in the intermittent fluid.

For Ψ being pressure head, θ being water content, K being the hydraulic conductivity of the porous medium, and f being a source term, the Richards hydrology model is the following:

\[
\begin{aligned}
\frac{\partial \theta}{\partial t} & (\Psi) - \text{div} \left( K(\theta(\Psi)) \nabla (\Psi + z) \right) = f(\theta(\Psi)), \quad (x, t) \in \Omega \times (0, T], \\
\theta(\Psi(x, 0)) &= \theta(\Psi_0(x)), \quad (x, t) \in \Omega \times \{0\}, \\
\Psi(x, t) &= \Psi_D(x, t), \quad (x, t) \in \Gamma_D \times (0, T], \\
K(\theta(\Psi)) \nabla (\Psi + z) \cdot \nu &= 0, \quad (x, t) \in \Gamma_N \times (0, T].
\end{aligned}
\]

Existence and uniqueness of solutions for Lipschitz continuous K and θ was proven by Alt and Luckhaus [1], after an application of the invertible Kirchoff transformation [4], which applies the variable transformation

\[ u = \int_{-\infty}^{\Psi} K(\theta(p)) \, dp, \]

to shift the nonlinearity in the divergence term to the lower order gravitational term. In practice however, we work with the original form (1.1), and so we will not discuss the Kirchoff Transformation further, other than to say that its invertibility carries the existence and uniqueness to the weak solution of (1.1).

2. Weak formulation

For the sake of simplicity, we consider the homogeneous Dirichlet problem, and assume that the initial pressure head profile Ψ_0(x) ∈ L^2(Ω). To solve this system, we multiply (1.1) by a test function \( v \in H^1_0(\Omega) \), integrate by parts, and use implicit Euler for the parabolic term to arrive at elliptic weak problems:

\[
(\theta^n, v) - \tau \left( \text{div} \left( K^n \nabla (\Psi^n + z) \right), v \right) = \tau(f^n, v) + (\theta^{n-1}, v), \quad v \in H^1_0(\Omega),
\]

1
where \((\cdot)^n\) is the value of the argument at timestep \(t_n = n\tau\). Here, \((\cdot, \cdot)\) represents the \(L^2(\Omega)\) inner product, with norm

\[
\|f\| := \left( \int_{\Omega} f^2 \, dx \right)^{1/2}.
\]

To solve the nonlinear elliptic problem at time \(t_n\), we linearize both nonlinearities, by considering the zeroth and first order Taylor expansions of \(K\) and \(\theta\), respectively. Omitting the time index \(n\), the linear weak problem to solve on each linearization step \(j\) for the correction \(\Psi_\epsilon\) is

\[
(\partial^j \Psi_\epsilon, v) - \tau (\text{div} (K^j \nabla \Psi_\epsilon), v) = R^j.
\]

This quasi-Newton method, popularized by Celia et al. \cite{3}, retains the mass conservation guaranteed by the continuous problem on the discrete level, and is widely used in the hydrological sciences \cite{2, 8, 5, 9, 7}. The main draw of this method (as opposed to the standard Newton-Raphson method) is that it generates symmetric linear problems to solve, that are typically better conditioned than the nonsymmetric Newton linear systems \cite{8, 5}. This typically comes at the cost of increased number of linear iterations required to converge, and a reduced robustness, particularly for problems modeling the infiltration of saturation fronts into initially dry media \cite{5, 8}. However, recent advancements in nonlinear solvers, such as those studied by Woodward et al. \cite{6}, have led to nonlinear solvers using modified Picard linear systems that are more robust than Newton’s method with respect to initial conditions and time step size, and on average, almost as fast to converge as Newton’s method.

3. Convergence properties

As part of my thesis work, we wanted to explore convergence properties of the modified Picard iteration, given sufficiently regular initial data, and Lipschitz continuous nonlinearities. Such conditions tend to arise when modeling infiltration into soils amenable to infiltration, such as sand and clay loam \cite{7}. A formal treatise on the convergence of this and other quasi-Newton methods for Richards’ equation is lacking in the literature; indeed, a comprehensive proof of convergence is hard to close, as the problem has multiple degeneracies, the most pressing being a degeneracy at the full saturation regime where the parabolic term vanishes entirely, rendering the problem fully elliptic. To simplify our analysis, we only consider the fully unsaturated regime.

Another problem occurs for initially dry media, where the Jacobian \(\partial \theta\) also tends asymptotically to 0, and so we choose a common assumption in the literature, that the pressure head is bounded away from the fully dry limit, which in turn enforces the properties

\[
0 < K_{\min} < K(\theta) < K_{\max},
\]

\[
0 < \partial \theta \leq L_{\theta}.
\]

Under these conditions, one can guarantee the convergence of the method, as when all of these conditions hold, the method can be seen as an adaptive form of the L-scheme proposed by Slodička \cite{10}. More precisely, the L-scheme as proposed in \cite{10} takes advantage of the monotonicity of \(\theta\) by replacing the \(\partial \theta\) term with an upper bound \(L > \partial \theta\):

\[
(L\Psi_\epsilon, v) - \tau (\text{div} (K^j \nabla \Psi_\epsilon), v) = R^j.
\]

This replacement effectively trades convergence speed for robustness, as Slodička is able to prove that for each fixed time step \(n\), the L-scheme forms a contraction, with convergence rate dependent on time step size, domain geometry, and \(K_{\min}\). These results carry straightforwardly for the modified Picard iteration in the case that we consider, as the same estimates carry through for each part of Slodička’s argument.
3.1. Slodička’s argument adjusted. Slodička’s proof of convergence as applied to the modified Picard iteration starts with showing that the iteration (2.2) forms a contraction, in the following sense:

**Lemma 3.1** ([10], Lemma 3.1). Let $K(\theta), \theta(\Psi)$ be Lipschitz continuous with respect to their arguments, with Lipschitz constants $L_k, L_\theta$ resp., $\Psi_0 = \Psi(x,0) \in L^2(\Omega)$, and the conditions (3.1) hold. Defining $\{\Psi_k\}_{k=1}^\infty$ as the linear iterates gotten by solving (2.2) and $\Psi$ the solution of the nonlinear form at time $t_n$ (2.1), then there exists a positive constant $\lambda = \lambda(\Omega, K_{\min})$ such that

$$
\|\Psi_k - \Psi\|^2 \leq \left(1 - \frac{\tau\lambda}{L + \tau\lambda}\right)^k \|\Psi_0 - \Psi\|^2,
$$

$$
\|\nabla(\Psi_k - \Psi)\|^2 \leq \frac{L + \tau\lambda}{\tau K_{\min}} \left(1 - \frac{\tau\lambda}{L + \tau\lambda}\right)^k \|\Psi_0 - \Psi\|^2,
$$

holds for all $k = 1, 2, \ldots$

This is proven by plugging in the error $\Psi - \Psi_k$ into (2.2), and rewriting the right hand side as a difference $h_k(\Psi) - h_k(\Psi_k)$, where

$$
h_k(s) := \theta(s) - \partial\theta(\Psi^{k-1})s, \quad s \in \mathbb{R}.
$$

Due to the Lipschitz nature of $\theta$, the linear coefficient $\partial\theta(\Psi^{k-1})$ can be bounded from above and below by $L$, and so the same error estimates as done by Slodička carry through.

Next, as this contraction holds, then for any time step $n$, there exists a maximal iteration index $\kappa$ such that

$$
\|\Psi^n - \Psi^n_{\kappa-1}\| \leq C_d \tau^d,
$$

where $d > 1$ and $C_d > 0$ are fixed constants. This, along with uniform bound $L > \partial\theta$ allows for the following a priori estimates, using the same calculations:

**Lemma 3.2** ([10], Lemmas 2.2 and 2.3). Let $K(\theta), \theta(\Psi)$ be Lipschitz continuous with respect to their arguments, with Lipschitz constants $L_k, L_\theta$ resp., $\Psi_0 = \Psi(x,0) \in L^2(\Omega)$, and the conditions (3.1) hold. Then the following a priori estimates hold for all $j = 1, 2, \ldots, n$:

$$
\tau \left(\sum_{i=1}^{j} \|\nabla\Psi_i\|^2 + \left\|\frac{\theta(\Psi_i) - \theta(\Psi_{i-1})}{\tau}\right\|_{-1}^2\right) \leq C,
$$

with $C$ independent of discretization parameters.

Here, the norm $\| \cdot \|_{-1}$ is the $H^{-1}(\Omega)$ norm, as is typically defined.

Finally, defining $\Psi_n(t)$ as the piecewise-constant in time function

$$
\Psi_n(t) = \begin{cases} 
\Psi_0, & t = 0, \\
\Psi^n, & t \in (t_{i-1}, t_i],
\end{cases}
$$

and piecewise-linear in time function

$$
w_n(t) = \begin{cases} 
\theta(\Psi_0), & t = 0, \\
\theta(\Psi_n(t_{i-1})) + (t - t_{i-1}) \frac{\theta(\Psi_n(t_i)) - \theta(\Psi_n(t_{i-1}))}{\tau}, & t \in (t_{i-1}, t_i],
\end{cases}
$$

and denoting $\bar{w}_n$ as the piecewise-constant in time equivalent of $w_n$, Slodička shows the weak convergence of $\Psi_n$ to the weak solution of (1.1), using the following theorems, which are compactness results as a consequence of the uniform estimates in Lemma 3.2:

**Theorem 1** ([10], Theorem 4.1). Let the assumptions of Lemma 3.1 be satisfied. Then,
(1) there exists a function \( w \in C(0, T; H^{-1}(\Omega)) \cap L_\infty(0, T; L_2(\Omega)) \) with \( \partial_t w \in L_2(0, T; H^{-1}(\Omega)) \) and a subsequence of \( \{ w_n \} \) for which
\[
\begin{align*}
w_n(t) &\to w(t) & \text{in} & \ C(0, T; H^{-1}(\Omega)), \\
w_n(t) &\rightharpoonup w(t) & \text{in} & \ L_2(\Omega), t \in [0, T], \\
\partial_t w_n &\rightharpoonup \partial_t w & \text{in} & \ L_2(0, T; H^{-1}(\Omega));
\end{align*}
\]

(2) there exists a function \( \Psi \in L_2(0, T; H^1_0(\Omega)) \) and a subsequence of \( \{ \Psi_n \} \) for which
\[
\Psi_n \rightharpoonup \Psi \quad \text{in} \quad L_2(0, T; H^1_0(\Omega)).
\]

**Theorem 2** ([10], Theorem 4.2). Let the assumptions of Lemma 3.1 be satisfied. Then,

(1) there exists a subsequence of \( \{ w_n \} \) for which
\[
\begin{align*}
w_n &\to w \quad \text{in} \quad L_2(\Omega \times [0, T]);
\end{align*}
\]

(2) \( w = \theta(\Psi) \), where \( \Psi \) is given by Theorem 4.1.

These two results, combined with Lemma 3.2 then yield weak convergence to the unique weak solution of (1.1):

**Theorem 3** ([10], Theorem 4.3). Let the assumptions of Lemma 3.1 be satisfied. Then the function \( \Psi \) given by Theorem 4.1 is a weak solution to (1.1).

The estimates as done by Slodička follow exactly the same for the modified Picard computations, after bounding the \( \partial_\theta(\Psi^{k-1}) \) terms with \( L \).

**References**