Generalized row-action methods for tomographic imaging

Martin S. Andersen · Per Christian Hansen

Received: 2 July 2013 / Accepted: 30 September 2013 / Published online: 1 November 2013
© Springer Science+Business Media New York 2013

Abstract Row-action methods play an important role in tomographic image reconstruction. Many such methods can be viewed as incremental gradient methods for minimizing a sum of a large number of convex functions, and despite their relatively poor global rate of convergence, these methods often exhibit fast initial convergence which is desirable in applications where a low-accuracy solution is acceptable. In this paper, we propose relaxed variants of a class of incremental proximal gradient methods, and these variants generalize many existing row-action methods for tomographic imaging. Moreover, they allow us to derive new incremental algorithms for tomographic imaging that incorporate different types of prior information via regularization. We demonstrate the efficacy of the approach with some numerical examples.

Keywords Incremental methods · Proximal methods · Inverse problems · Regularization · Tomographic imaging

Mathematics Subject Classifications (2010) 65R32 · 90C52 · 65K05
1 Introduction

Tomographic imaging is an indispensable non-invasive measurement technique for diagnostics, exploration, analysis, and design; see [4, 25, 33] and the references therein. Discretizations of tomographic imaging problems often lead to large sparse systems of linear equations with noisy data:

\[ Ax \approx b, \quad A \in \mathbb{R}^{m \times n}. \tag{1} \]

Here the vector \( x \) represents the unknown image, the vector \( b \) is the given (usually inaccurate/noisy) data, and the matrix \( A \) models the forward problem. There are no restrictions on the dimensions of \( A \), and both over- and underdetermined systems arise in applications, depending on the amount of data generated in a given experiment.

Iterative algorithms are often well-suited for solving the large-scale problem (1), and several classes of methods have emerged [21, 26]. They all produce regularized solutions that approximate the exact and unknown solution image without being too sensitive to the perturbation of the data.

This work focuses on a specific class of so-called row-action methods, the basic form of which is known as Kaczmarz’s method or ART (algebraic reconstruction technique) [23, 28]. These methods have been used for several decades as the core computational routines for tomographic imaging, and they are recognized for often having fast initial convergence towards the desired image. An important advantage of these methods is that they access the matrix \( A \) one row—or one block—at a time, thus making the methods well suited for modern computer architectures.

Several extensions of the classical (block) ART methods have been proposed with the goal of improving certain characteristics of the reconstructed images. Of particular interest is the use of total variation (TV) regularization as a way to better preserve edges and detail in the image. For example, Censor, Davidi, and Herman et al. [16] developed a so-called “perturbation resilient” framework to incorporate TV regularization into the ART iterations, while Sidky and Pan [39] proposed a hybrid algorithm where ART is combined with the steepest descent method, also to incorporate TV regularization.

The main goal of this paper is to provide a theoretical and algorithmic framework for studying and generalizing the ART methods. The cornerstone of our approach is an interpretation of ART as a so-called incremental proximal gradient method for convex optimization. This allows us to generalize the method (e.g. with the TV regularization term) in a rigorous way—thus avoiding the heuristic arguments sometimes found in applications.

The main contribution of this paper is twofold: (i) we propose a generalization of the incremental proximal gradient framework of Bertsekas [7, 8] that includes a relaxation parameter, and (ii) using this framework, we propose a class of generalized row-action methods that allows us to incorporate different kinds of prior information in the reconstruction problem via regularization.

The paper is organized as follows. In Section 2, we discuss incremental methods and proximal methods for convex optimization, and in Section 3, we present
two relaxed incremental proximal gradient methods. We discuss some connections between existing row-action methods and the relaxed incremental proximal gradient framework in Section 4, and in Section 5, we consider generalized row-action methods for data fitting with a regularization term. We present some numerical results in Section 6, and we conclude the paper in Section 7.

**Notation** The $i$th row of $A$ is denoted by $a_i^T$, and $A^\dagger$ denotes the Moore–Penrose pseudoinverse of $A$. Given a convex function $f : \mathbb{R}^n \to \mathbb{R} \cup \{\pm \infty\}$, we denote by $\text{dom} f = \{x \in \mathbb{R}^n \mid f(x) < \infty\}$ the effective domain of $f$. Finally, $P_C(x)$ denotes the Euclidean projection of $x \in \mathbb{R}^n$ on a closed convex set $C$ of $\mathbb{R}^n$, and $\text{dist}(x, C) \equiv \|x - P_C(x)\|_2$ denotes is the Euclidean distance from $x$ to the set $C$.

## 2 The optimization framework

Many reconstruction problems in tomographic imaging can be expressed as a constrained convex optimization problem with an objective function that is given as a sum of $m$ convex functions, i.e.,

$$\begin{aligned}
\text{minimize} & \quad f(x) \equiv \sum_{i=1}^{m} f_i(x) \\
\text{subject to} & \quad x \in C.
\end{aligned}$$

(2)

Here $x \in \mathbb{R}^n$ is the optimization variable, $f_i : \mathbb{R}^n \to \mathbb{R}$, and $C$ is a closed convex subset of $\mathbb{R}^n$. The functions $f_i$ typically represent data-fidelity terms, such as squared residuals and one or more regularization terms that incorporate prior knowledge. The set $C$ may represent bounds on the components of $x$; in this paper we will simply assume that the projection operator $P_C(\cdot)$ associated with the set $C$ is cheap to evaluate.

If the objective function in (2) is differentiable with a Lipschitz continuous gradient, the problem can be solved using an accelerated gradient projection method. If $x_k$ denotes the $k$th iterate and $f^*$ is the minimum, then the error $f(x_k) - f^*$ is $O\left(1/k^2\right)$; see, e.g., [34]. For problems with a nonsmooth objective function that is Lipschitz continuous on a bounded set, the error bound is $O\left(1/\sqrt{k}\right)$, and this can be achieved using a projected subgradient method with a diminishing step-size rule. This error bound can often be improved by exploiting problem structure. For example, the accelerated proximal gradient method of Beck and Teboulle [2] splits the objective function into a smooth term and a nonsmooth term, and this method achieves the same error bound as the accelerated methods for smooth optimization, namely $O\left(1/k^2\right)$.

Before we turn to the main subject of the paper in Section 3, relaxed incremental proximal gradient methods, we briefly review some necessary material.

### 2.1 Incremental gradient methods

When the objective function in (2) is comprised of a very large number of functions, the cost of computing the gradient (or a subgradient) may be very high. To avoid
computing the full gradient, incremental gradient methods use only the gradient of a single component of the objective function at iteration $k$, i.e.,

$$x_{k+1} = \mathcal{P}_C(x_k - t_k \nabla f_{i_k}(x_k)),$$

where $i_k \in \{1, \ldots, m\}$ is the index of the component used for the update at iteration $k$. The index $i_k$ is commonly chosen either in a cyclic manner (e.g., $i_k = (k \mod m) + 1$) or drawn uniformly at random; another possibility is to combine the cyclic rule with randomization by shuffling the order of the indices at the beginning of each cycle, and empirical evidence suggests that this works very well in practice [8, 35].

Incremental methods typically have a very slow asymptotic rate of convergence, and like subgradient methods they require a diminishing step-size rule to ensure convergence. In tomographic applications, however, we are more interested in the initial rate of convergence (and the associated semi-convergence [33]), and this can be very fast for incremental methods compared to their nonincremental counterparts [5, 6, 8, 20]. There are also several examples of hybrid methods that gradually transition from an incremental method to a full gradient method in order to combine the fast initial convergence of the incremental method and the asymptotic rate of the full gradient method; see, e.g., [5, 9, 22] and references therein.

2.2 Proximal methods

Given a closed convex function $f : \mathbb{R}^n \to \mathbb{R}$, the proximal operator $\text{prox}_f(x) : \mathbb{R}^n \to \mathbb{R}^n$ associated with $f$ is defined as follows [31]

$$\text{prox}_f(x) = \arg \min_{u \in \mathbb{R}^n} \left\{ f(u) + 1/2\|u - x\|^2 \right\}. \quad (3)$$

The first-order optimality condition associated with the minimization in (3) can be expressed as $x - u \in \partial f(u)$, where $\partial f(u)$ denotes the subdifferential of $f$ at $u \in \text{dom} f$, defined as

$$\partial f(u) = \left\{ w \in \mathbb{R}^n \mid f(y) \geq f(u) + w^T(y - u) \right\}. \quad (4)$$

In particular, if $f$ is differentiable at $u$, then $\partial f(u)$ is the singleton $\{ \nabla f(u) \}$ where $\nabla f(u)$ denotes the gradient of $f$ at $u$. It follows that if $x$ is a fixed-point of $\text{prox}_f(x)$ (i.e., if $x = \text{prox}_f(x)$), then $0 \in \partial f(x)$ and hence $x$ is a minimizer of $f$. In other words, minimizing $f$ is equivalent to finding a fixed-point of $\text{prox}_f(x)$.

The proximal operator associated with the indicator function of a closed convex set $C$ of $\mathbb{R}^n$, defined as

$$I_C(x) = \begin{cases} 0 & x \in C \\ \infty & \text{otherwise}, \end{cases}$$

is simply the Euclidean projection of $x$ on $C$, i.e., $\text{prox}_{I_C}(x) = \arg \min_{x \in \mathbb{R}^n} \|u - x\|^2 = \mathcal{P}_C(x)$. It is therefore natural to view the proximal operator associated with a closed convex function $f$ as a generalized projection operator.

The proximal point method, proposed by Martinet [29, 30] in the early 1970s and further studied by Rockafellar [37], is a method for solving monotone inclusion problems of the form $0 \in T(x)$ where $T$ is a maximal monotone operator. Since

© Springer
the subdifferential operator $\partial f$ associated with a closed convex function $f$ is maximal monotone \[36\], the proximal point algorithm can be used to solve the inclusion problem $0 \in \partial f(x)$. For this problem, the proximal point algorithm can be written as

$$x_{k+1} = \text{prox}_{t_k f}(x_k) = \arg \min_{u \in \mathbb{R}^n} \left\{ t_k f(u) + 1/2\|u - x_k\|^2 \right\}, \quad (5)$$

where $\{t_k\}$ is a sequence of positive parameters. It follows from the optimality condition $(x_k - u)/t_k \in \partial f(u)$ associated with the minimization in (5) that the proximal point algorithm can be expressed as

$$x_{k+1} = x_k - t_k \tilde{\nabla} f(x_{k+1}),$$

where $\tilde{\nabla} f(x_{k+1}) = (x_k - x_{k+1})/t_k$ is a subgradient that belongs to the subdifferential $\partial f(x_{k+1})$, and $t_k$ is an implicit step-size parameter. Thus, the proximal point method can be viewed as an implicit (sub)gradient method, and unlike the standard gradient method, it converges for any positive sequence $\{t_k\}$, provided that a minimum exists.

2.3 Incremental proximal gradient methods

The incremental proximal gradient methods of Bertsekas \[8\] seek to minimize a sum $f(x) = \sum_{i=1}^m f_i(x)$ over a closed convex subset $C$ of $\mathbb{R}^n$, and each $f_i$ is a sum of two convex functions $f_i(x) = g_i(x) + h_i(x)$ with $g_i : \mathbb{R}^n \to \mathbb{R}$ and $h_i : \mathbb{R}^n \to \mathbb{R}$, i.e.,

$$\text{minimize } f(x) = \sum_{i=1}^m (g_i(x) + h_i(x))$$

subject to $x \in C$. \[6\]

We will assume that each of the functions $g_i$ possesses “favorable structure” so that the proximal minimization $\text{prox}_{\hat{g}_i}(x)$ is easy to solve or has a closed-form solution.

Given the $k$th iterate $x_k$, an index $i_k \in \{1, \ldots, m\}$, and a step size $t_k > 0$, Bertsekas’ first incremental proximal method, which we denote IPG1, computes $x_{k+1}$ as follows

(IPG1) \[7a\] $z_k = \text{prox}_{t_k g_{i_k}}(x_k)$

$$x_{k+1} = P_C \left( z_k - t_k \tilde{\nabla} h_{i_k}(z_k) \right). \quad (7b)$$

Here $\tilde{\nabla} h_{i_k}(z_k)$ denotes a subgradient of $h_{i_k}$ at $z_k$, i.e., $\tilde{\nabla} h_{i_k}(z_k) \in \partial h_{i_k}(z_k)$. From the optimality conditions associated with (7a) and since $z_k$ is unique, we have that $z_k = x_k - t_k \tilde{\nabla} g_{i_k}(z_k)$ for some $\tilde{\nabla} g_{i_k}(z_k)$ in the subdifferential of $g_{i_k}$ at $z_k$. Substituting this expression for $z_k$ in (7b), we obtain the following equivalent expression for $x_{k+1}$:

$$x_{k+1} = P_C \left( x_k - t_k \tilde{\nabla} f_{i_k}(z_k) \right). \quad (8)$$

Thus, IPG1 can be viewed as an incremental extragradient-like method where $x_{k+1}$ is obtained by first computing a “predictor” $z_k$, followed by a subgradient step based on a subgradient of $f_{i_k}$, evaluated at the predictor $z_k$ instead of at the current iterate $x_k$.

The second incremental proximal method of Bertsekas, which we will refer to as IPG2, can be expressed as the iteration

(IPG2) \[9a\] $z_k = x_k - t_k \tilde{\nabla} h_{i_k}(x_k)$

$$x_{k+1} = \text{prox}_{t_k \hat{g}_{i_k}}(z_k). \quad (9b)$$
where $\tilde{g}_{ik}(x) = g_{ik}(x) + I_C(x)$. If we substitute (9a) for $z_k$ in (9b), we obtain the equivalent formulation

$$
x_{k+1} = \arg\min_{x \in \mathcal{C}} \left\{ g_{ik}(x) + \tilde{\nabla} h_{ik}(x_k)^T x + \frac{1}{2t_k} \|x - x_k\|^2 \right\}.
$$

(10)

As pointed out in [8], IPG2 can be viewed as an incremental version of the iterative shrinkage/thresholding algorithm [12, 15]. Alternatively, IPG2 can be viewed as an incremental proximal algorithm with partial linearization of the component function $f_i$ (i.e., only $h_i$ is linearized).

Both of the methods IPG1 and IPG2 reduce to the same incremental (sub)gradient algorithm if $g_i(x) = 0$ for all $i$. However, if $h_i(x) = 0$ for all $i$, we obtain two slightly different incremental proximal methods in general: IPG1 involves an unconstrained optimization which is followed by an explicit projection on $\mathcal{C}$ whereas IPG2 includes the constraint $x \in \mathcal{C}$ in the minimization (10).

### 3 Relaxed incremental proximal gradient methods

It is well-known that the performance of many classical row-action methods for tomographic imaging depend strongly on a relaxation parameter. Motivated by this, we now propose relaxed variants of the incremental proximal gradient methods.

#### 3.1 A modified IPG2 method

In some applications, the constraint $x \in \mathcal{C}$ in (10) prohibits a closed-form solution or cheap computation of the solution to the proximal minimization, and in such applications, IPG1 may be more suitable than IPG2. To overcome this limitation of IPG2, we propose a modified variant of IPG2 which omits the constraint $x \in \mathcal{C}$ from the minimization (10) and instead adds a projection step, i.e.,

$$
z_k = \arg\min_{x \in \mathbb{R}^n} \left\{ g_{ik}(x) + \tilde{\nabla} h_{ik}(x_k)^T x + \frac{1}{2t_k} \|x - x_k\|^2 \right\},
$$

(11a)

$$
x_{k+1} = \mathcal{P}_\mathcal{C}(z_k),
$$

(11b)

or equivalently, if we combine the two steps,

$$
x_{k+1} = \mathcal{P}_\mathcal{C} \left( \text{prox}_{t_k g_{ik}}(x_k - t_k \tilde{\nabla} h_{ik}(x_k)) \right).
$$

Note that this modified version of IPG2 is equivalent to IPG1 when $h_i(x) = 0$ for all $i$.

#### 3.2 The R-IPG1 and R-IPG2 methods

We are now ready to propose relaxed variants of IPG1 and the modified IPG2 in (11). The relaxed variant of IPG1, which we will call R-IPG1, depends on a relaxation parameter $\rho \in (0, 2)$, and it is defined as the iteration

$$
(R\text{-IPG1}) \quad w_k = \text{prox}_{t_k g_{ik}}(x_k)
$$

(12a)
\[ z_k = w_k - t_k \tilde{\nabla} h_{i_k} (w_k) \quad (12b) \]
\[ x_{k+1} = P_{\mathcal{C}} (\rho z_k + (1 - \rho)x_k). \quad (12c) \]

Similarly, R-IPG2 refers to the relaxed variant of (11), and it is defined as
\[
(R\text{-IPG2}) \quad w_k = x_k - t_k \tilde{\nabla} h_{i_k} (x_k) \quad (13a)
\]
\[ z_k = \text{prox}_{t_k g_{i_k}} (w_k) \quad (13b) \]
\[ x_{k+1} = P_{\mathcal{C}} (\rho z_k + (1 - \rho)x_k). \quad (13c) \]

Notice that the relaxed algorithms (12) and (13) are very similar, and they differ only in the order of the first two updates at each iteration.

**Remark 1** It is easy to verify that the two relaxed methods produce the exact same sequence \(\{x_k\}\) if either \(h_i(x) = 0\) or \(g_i(x) = 0\) for all \(i\). In the latter case, both R-IPG1 and R-IPG2 reduce to a projected (sub)gradient method with step size \(\rho t_k\), and this implies that the relaxation parameter is redundant when \(g_i(x) = 0\) for all \(i\).

### 3.3 Convergence results

We now address the convergence properties of R-IPG1 and R-IPG2 using cyclic control. Following the exposition in [8], we will make the following assumptions about the functions \(g_i\) and \(h_i\) and their (sub)gradients.

**Assumption 1** (R-IPG1) There exists a constant \(c\) such that for all \(k\),
\[
\max \{ \| \tilde{\nabla} g_{i_k} (w_k) \|_2, \| \tilde{\nabla} h_{i_k} (w_k) \|_2 \} \leq c \quad (14)
\]
and for all \(k\) that mark the beginning of a cycle, we have for all \(j = 1, \ldots, m\),
\[
\max \{ g_j(x_k) - g_j(w_{k+j-1}), h_j(x_k) - h_j(w_{k+j-1}) \} \leq c \| x_k - w_{k+j-1} \|_2. \quad (15)
\]

**Assumption 2** (R-IPG2) There exists a constant \(c\) such that for all \(k\),
\[
\max \{ \| \tilde{\nabla} g_{i_k} (z_k) \|_2, \| \tilde{\nabla} h_{i_k} (x_k) \|_2 \} \leq c \quad (16)
\]
and for all \(k\) that mark the beginning of a cycle, we have for all \(j = 1, \ldots, m\),
\[
\max \{ g_j(x_k) - g_j(x_{k+j-1}), h_j(x_k) - h_j(x_{k+j-1}) \} \leq c \| x_k - x_{k+j-1} \|_2 \quad (17)
\]
\[
g_j(x_{k+j-1}) - g_j(z_{k+j-1}) \leq c \| x_{k+j-1} - z_{k+j-1} \|_2. \quad (18)
\]

**Remark 2** Assumptions 1 and 2 are satisfied if, for example, all \(g_i\) and \(h_i\) are Lipschitz continuous on \(\mathbb{R}^n\), or if the sequences \(\{x_k\}\) and \(\{w_k\}\) (in the case of R-IPG1) or \(\{x_k\}\) and \(\{z_k\}\) (in case of R-IPG2) are bounded. See [8] for further details.

A key component of the convergence analysis is the following generalization of Proposition 3 in [8].
Proposition 1 Let \( \{x_k\} \) be a sequence generated by either (12) or (13) with the index \( i_k \) chosen according to the cyclic rule \( i_k = (k \mod m) + 1 \). Then, given a point \( y \in \mathcal{C} \) and a relaxation parameter \( \rho \in [\delta, 2 - \delta] \) for some \( \delta > 0 \),
\[
\|x_{k+m} - y\|_2^2 \leq \|x_k - y\|_2^2 - 2\rho t_k (f(x_k) - f(y)) + \beta \rho^2 t_k^2 m^2 c^2
\]  
where
\[
\beta = \begin{cases} 
4 + \frac{1-\rho+\alpha}{\rho m} & \text{for R-IPG1 (12)} \\
4 + \frac{4(1-\rho)+\alpha}{\rho m} & \text{for R-IPG2 (13)},
\end{cases}
\]
in which \( \alpha \) is a constant defined as
\[
\alpha = \begin{cases} 
1/(2 - \rho) & \delta \leq \rho \leq 3/2 \\
4(1 - \rho) & 3/2 < \rho \leq 2 - \delta.
\end{cases}
\]

Proof See Appendix A.

Remark 3 If we let \( \rho = 1 \) in Proposition 1, then we obtain \( \beta = 4 + 1/m \) for both of the relaxed methods. The constant for IPG2 derived in [8] is \( \beta = 4 + 5/m \); this discrepancy arises because of an approximation \( m^2 \approx m^2 - m \) in the proof in [8], and without this approximation we obtain \( \beta = 4 + 1/m \) for both IPG1 and IPG2. Figure 1 shows the constant \( \beta \) as a function of \( \rho \) for both R-IPG1 and R-IPG2 and different values of \( m \).

The following proposition summarizes the main convergence results for problems where \( f(x) \) is bounded below and using cyclic control.

Proposition 2 Let \( \{x_k\} \) be a sequence generated by (12) or (13), and suppose \( f(x) \) is bounded below (\( f^* > -\infty \)). Then, using cyclic control we have an error bound for constant step size \( t_k = t \)
\[
\liminf_{k \to \infty} f(x_k) = f^* + \frac{\rho t \beta m^2 c^2}{2}
\]
and exact convergence for a diminishing step-size rule that satisfies \( \sum_{k=1}^{\infty} t_k = \infty \) and \( \lim_{k \to \infty} t_k = 0 \), i.e.,
\[
\liminf_{k \to \infty} f(x_k) = f^*.
\]

Proof The bound (19) has the exact same form as the bound in Proposition 3 in [8] if we define a scaled parameter \( \tilde{t}_k = \rho t_k \). Since the convergence analysis in [8] is based on this bound, it also holds for the relaxed variants of the methods; see Propositions 4 and 6 in [8].

Remark 4 It is also possible to obtain similar bounds (in expectation) for R-IPG1 and R-IPG2 with randomized control. However, since the analysis is nearly identical to that in [8], we omit the details for the sake of brevity. Further details and efficiency estimates can be found in [8].
The error bound for the above methods is $O(mc/\sqrt{\ell})$ where $\ell$ is the number of cycles. We remind the reader that despite this poor global error bound, the incremental methods often have fast initial rate of convergence and may outperform nonincremental methods when low-accuracy is acceptable.

4 ART within the R-IPG framework

This section shows an important application of the algorithmic framework introduced in the previous section. In particular, we demonstrate how specialized variants of the relaxed algorithms lead to the well-known ART method and variants of this method.

4.1 Relaxed ART methods

ART can be viewed both as an incremental gradient method and as an incremental proximal method. Specifically, if we let $g_i(x) = 0$ and $h_i = 1/2 \left( a_i^T x - b_i \right)^2 / \|a_i\|^2_2$, both R-IPG1 and R-IPG2 result in the incremental gradient iteration

$$x_{k+1} = \mathcal{P}_C \left( x_k - \rho t_k a_i \frac{a_i^T x_k - b_i}{\|a_i\|^2_2} \right),$$

which is equivalent to ART with relaxation parameter $\rho$ if we let $t_k = 1$. Similarly, if we let $g_i(x) = I_{H_i}(x)$ and $h_i = 0$ where $H_i = \{ x \in \mathbb{R}^n | a_i^T x - b_i = 0 \}$, then both R-IPG1 and R-IPG2 result in the following algorithm

$$x_{k+1} = \mathcal{P}_C \left( \rho \mathcal{P}_{H_i} (x_k) + (1 - \rho) x_k \right),$$

where the projection of $x_k$ on $H_i$ is $\mathcal{P}_{H_i} (x_k) = x_k - a_i \frac{a_i^T x_k - b_i}{\|a_i\|^2_2}$.

When we insert this relation into (21), we once again obtain ART with relaxation parameter $\rho$.

Note that although the iteration (21) is an incremental proximal algorithm, the choice $g_i(x) = I_{H_i}(x)$ does not satisfy Assumptions 1 and 2. In fact, the corresponding problem is a convex feasibility problem that may or may not be feasible. ART, however, is known to converge to the minimum norm solution if the system $Ax = b$
is consistent (i.e., the feasibility problem is feasible), and otherwise ART converges to a weighted least-squares solution provided that a diminishing step-size sequence is used [11, 27]. Similarly, using Proposition 2, it follows that the iteration (20) converges to a weighted least-squares solution when a diminishing step-size sequence is used.

An alternative to the choice \( g_i(x) = I_{H_i}(x) \) is to define \( g_i(x) = \text{dist}(x, H_i) \) or \( g_i(x) = \text{dist}(x, H_i)^2 \), and as we will see in the next section, this gives rise to damped ART-like algorithms.

### 4.2 Damped ART

It is an interesting and useful fact that there are many other possible choices of \( g_i \) and \( h_i \) that lead to convergent incremental methods that are similar to ART. For example, if we let \( g_i(x) = \frac{1}{2} \left( a_i^T x - b_i \right)^2 \) and \( h_i(x) = 0 \), we obtain the following incremental proximal method

\[
    x_{k+1} = \mathcal{P}_C \left( x_k - \rho \frac{a_{ik} \left( a_{ik}^T x_k - b_{ik} \right)}{\|a_{ik}\|^2_2 + t_k^{-1}} \right).
\]

This iteration can be viewed as a damped ART method where \( t_k \) determines the damping at step \( k \). Large values of \( t_k \) correspond to a small amount of damping, and it is easy to verify that in the limit, if we let \( t_k \to \infty \), the iteration is equivalent to ART. This variant of ART is useful when some rows have very small but nonzero norm, in which case the damping helps to suppress noise amplification—we illustrate this with an example in Section 6.1.

It is also possible to derive generalized block methods based on the relaxed incremental proximal gradient methods. Here we consider block iterative methods for minimizing \( \|Ax - b\|^2_2 \). Suppose we partition \( A \) and \( b \) into \( p \) blocks of rows where \( A_i \in \mathbb{R}^{m_i \times n} \) denotes the \( i \)th block of \( A \) and \( b_i \in \mathbb{R}^{m_i} \) denotes the \( i \)th block of \( b \). A variant of Elfving’s block-Kaczmarz method [18] then follows from iteration (21) if we let \( B_i = \{ x \in \mathbb{R}^n | A_i x = b_i \} \). Since the projection of a point \( x \) onto \( B_i \) can be expressed as \( \mathcal{P}_{B_i}(x) = x - A_i^\dagger (A_i x - b_i) \), we can express the block Kaczmarz method as

\[
    x_{k+1} = x_k - \rho A_{ik}^\dagger \left( A_{ik} x - b_{ik} \right).
\]

Notice that like Kaczmarz’s method, the block Kaczmarz method does not include the parameter \( t_k \). If we instead let \( g_i(x) = \frac{1}{2} \|A_i x - b_i\|^2_2 \), we obtain the following proximal operator

\[
    \text{prox}_{t_k g_i}(x) = \left( I + t_k A_{ik}^T A_{ik} \right)^{-1} \left( x + t_k A_{ik}^T b_{ik} \right)
\]

\[
    = x - M_{ik} \left( A_{ik} x - b_{ik} \right),
\]

where \( M_{ik} = A_{ik}^T \left( A_{ik} A_{ik}^T + t_k^{-1} I \right)^{-1} \). From the limit definition [1]

\[
    A^\dagger = \lim_{\delta \to 0} \left( A^T A + \delta^2 I \right)^{-1},
\]

\[
    A = \lim_{\delta \to 0} A^T \left( A A^T + \delta^2 I \right)^{-1},
\]

\( \rho \text{ Springer} \)
we immediately see that $M_{ik} \rightarrow A_{ik}^\dagger$ as $t_k \rightarrow \infty$. We can therefore interpret $M_{ik}$ as a regularized or damped pseudoinverse of $A_{ik}$, and hence the resulting incremental method is a block variant of (22).

We obtain yet another damped ART-like algorithm if we let $h_i(x) = 0$ and either $g_i(x) = \text{dist}(x, \mathcal{H}_i)$ or $g_i(x) = \frac{1}{2}\text{dist}(x, \mathcal{H}_i)^2$. For example, if we let $g_i(x) = \frac{1}{2}\text{dist}(x, \mathcal{H}_i)^2$, the proximal operator associated with $g_{ik}$ is given by

$$\text{prox}_{t_k g_{ik}}(u) = (1 - \theta_k)u + \theta_k P_{\mathcal{H}_{ik}}(u),$$

where $\theta_k = \min \left(1, t_k \|a_{ik}\|_2/\|a_{ik}^Tu - b_{ik}\|\right)$; see, e.g., [14]. The resulting relaxed incremental proximal iteration is of the form

$$x_{k+1} = \begin{cases} 
\mathcal{P}_{\mathcal{C}} \left( x_k - \rho \left( a_{ik} \frac{a_{ik}^Tx_k - b_{ik}}{\|a_{ik}\|^2_2} \right) \right) & \text{if } a_{ik}^Tx_k - b_{ik} < t_k \|a_{ik}\|_2 \\
\mathcal{P}_{\mathcal{C}} \left( x_k - \rho t_k \frac{a_{ik}}{\|a_{ik}\|_2} \text{sgn} \left( a_{ik}^Tx_k - b_{ik} \right) \right) & \text{otherwise} 
\end{cases}$$

(26)

if $\|a_{ik}\|_2 > 0$, and otherwise $x_{k+1} = x_k$.

4.3 Damped ART for robust regression

It is well-known that the least squares objective $\|Ax - b\|_2^2$ yields a maximum a posteriori estimate when the noise is Gaussian. However, $\ell_2$ data fitting is sensitive to outliers. A more robust criterion is the $\ell_1$ norm objective $\|Ax - b\|_1$ (which is also known as linear least absolute value regression), and this yields a maximum a posteriori estimate when the noise follows a Laplace distribution [17]. Minimizing the $\ell_1$ norm of the residuals instead of the squared $\ell_2$ norm leads to another damped ART-like algorithm. Specifically, if we let $g_i(x) = |a_i^Tx - b_i|$ and $h_i(x) = 0$, then both R-IPG1 and R-IPG2 lead to the following update

$$x_{k+1} = \begin{cases} 
\mathcal{P}_{\mathcal{C}} \left( x_k - \rho \left( a_{ik} \frac{a_{ik}^Tx_k - b_{ik}}{\|a_{ik}\|^2_2} \right) \right) & \text{if } a_{ik}^Tx_k - b_{ik} < t_k \|a_{ik}\|_2 \\
\mathcal{P}_{\mathcal{C}} \left( x_k - \rho t_k a_{ik} \text{sgn} \left( a_{ik}^Tx_k - b_{ik} \right) \right) & \text{otherwise} 
\end{cases}$$

(27)

To see this, consider the proximal operator

$$\text{prox}_{t_k g_{ik}}(x_k) = \arg \min_{u \in \mathbb{R}} \left\{ t_k \left| a_{ik}^Tu - b_{ik} \right| + \frac{1}{2} \|u - x_k\|^2_2 \right\}.$$
Clearly, $\text{prox}_{\iota_k g_k}(x_k) = x_k$ if either $a_{ik} = 0$ or $a_{ik}^T x_k = b_{ik}$, and otherwise the minimizer $u^*$ is attained on the line segment between $x_k$ and its projection on the affine subspace $\{ x | a_{ik}^T x = b_{ik} \}$, i.e.,

$$u^* = x_k - \theta^* a_{ik} \left( a_{ik}^T x_k - b_{ik} \right) / \| a_{ik} \|^2_2 \tag{28}$$

for some $\theta^* \in (0, 1]$. Thus, using this parameterization of $u^*$, we may evaluate the proximal operator by computing

$$\theta^* = \arg \min_{\theta \in [0, 1]} \left\{ t_k (1 - \theta) \left| a_{ik}^T x_k - b_{ik} \right| + \theta^2 / 2 \left( a_{ik}^T x_k - b_{ik} \right)^2 / \| a_{ik} \|^2_2 \right\}$$

$$= \min \left( 1, t_k \| a_{ik} \|^2_2 / \left| a_{ik}^T x_k - b_{ik} \right| \right)$$

and the minimizer $u^* = \text{prox}_{\iota_k g_k}(x_k)$ then follows from (28). Using this result in R-IPG1 or R-IPG2, we obtain the iteration (27). This is very similar to the iteration (26), the step-size rule is based on $\| a_{ik} \|^2_2$ instead of $\| a_{ik} \|^2_2$.

The update (27) is simply a (relaxed) projection if the magnitude of the residual $\left| a_{ik}^T x_k - b_{ik} \right|$ is sufficiently small. On the other hand, if $\left| a_{ik}^T x_k - b_{ik} \right|$ is sufficiently large, the step will be damped, and the parameter $t_k$ governs the damping. Notice that like the method (22), the method (27) also reduces to ART if $t_k$ is sufficiently large.

The Huber penalty $\phi_{\mu}(t)$, which is defined as

$$\phi_{\mu}(t) = \begin{cases} t^2 / (2\mu) & |t| < \mu \\ |t| - \mu / 2 & |t| \geq \mu \end{cases}$$

where $\mu \geq 0$ is a parameter, can be viewed as a combination of the $\ell_1$ and $\ell_2$ norms. If we define $g_i(x) = \phi_{\mu} \left( a_i^T x - b_i \right)$ and $h_i(x) = 0$ in R-IPG1 or R-IPG2, we obtain the following iteration

$$x_{k+1} = \begin{cases} \mathcal{P}_C \left( x_k - \rho_{t_k} \frac{a_{ik} \left( a_{ik}^T x_k - b_{ik} \right)}{\mu / t_k + \| a_{ik} \|^2_2} \right) & \left| a_{ik}^T x_k - b_{ik} \right| < \mu + t_k \| a_{ik} \|^2_2 \\ \mathcal{P}_C \left( x_k - \rho_{t_k} a_{ik} \| a_{ik} \|^2_2 \text{sgn} \left( a_{ik}^T x_k - b_{ik} \right) \right) & \text{otherwise.} \end{cases} \tag{29}$$

Note that this reduces to the algorithm (27) for $\ell_1$ norm minimization when $\mu = 0$.

### 5 Generalized row-action methods with regularization

We now turn to constrained regularized least-squares problems of the form

$$\text{minimize } f(x) = 1/2 \sum_{i=1}^m \left( a_i^T x - b_i \right)^2 + \lambda \psi(x) \tag{30}$$

subject to $x \in \mathcal{C}$,

where $\lambda > 0$ is a regularization parameter. We will assume that the regularization function $\psi(x)$ is convex (but not necessarily smooth). Notice that this problem is of the form (6) if, for example, we let $g_i(x) = 1/2 \left( a_i^T x - b_i \right)^2$ and $h_i(x) = \lambda / m \psi(x)$.
There are obviously many other ways to express (30) as a problem of the form (6), and these give rise to a family of relaxed incremental proximal gradient algorithms for the problem (30). Note that although the resulting algorithms may appear to be somewhat similar, they may behave very differently in practice.

A straightforward way to construct an incremental method for the regularized least-squares problem (30) is to define \( m \) components

\[
g_i(x) = \frac{1}{2} \left( a_i^T x - b_i \right)^2, \quad h_i(x) = \frac{\lambda}{m} \psi(x), \quad i = 1, \ldots, m.
\]

This choice results in algorithms that alternate between a small (sub)gradient step for the regularization term, a damped projection on a hyperplane defined by one of the equations \( a_i^T x - b_i = 0 \), and a projection on \( \mathcal{C} \).

We obtain a different pair of algorithms if we instead define \( m + s \) components

\[
g_i(x) = \frac{1}{2} \left( a_i^T x - b_i \right)^2, \quad h_i(x) = 0, \quad i = 1, \ldots, m, \tag{32a}
\]

\[
g_{m+i}(x) = \frac{\lambda}{s} \psi(x), \quad h_{m+i}(x) = 0, \quad i = 1, \ldots, s, \tag{32b}
\]

or alternatively, instead of (32b),

\[
g_{m+i}(x) = 0, \quad h_{m+i}(x) = \frac{\lambda}{s} \psi(x), \quad i = 1, \ldots, s. \tag{32c}
\]

With cyclic component selection, this results in ART-like algorithms where each cycle consists of \( m \) damped projections, followed by either \( s \) proximal steps associated (32b) or \( s \) (sub)gradient steps associated with (32c).

These algorithms are similar to the “superiorization method” of Censor et al. [10] which is also an ART-like method where each cycle consists of a complete ART-cycle, followed by a “correction step” which is referred to as a perturbation. The correction step can be a (sub)gradient step, and if the set \( D = \{ x \mid x \in \mathcal{C}, Ax = b \} \) is nonempty, then the superiorization method converges to a point in \( D \) provided that the norm of the correction at iteration \( k \) goes to zero as \( k \to \infty \). If the correction is a negative subgradient of a regularization function \( \psi(x) \), the method tends to converge to points in \( D \) for which the regularization term \( \psi(x) \) is small compared to what can be achieved with plain ART. In contrast, our approach leads to methods that always converge to a solution of the regularized least-squares problem (30), provided that a diminishing step-size sequence is used.

5.1 Least-squares with total variation regularization

Total Variation (TV) regularization [38] is popular in imaging because of its ability to suppress noise while preserving edges. In the discrete setting, the TV seminorm of a vector representation \( x \in \mathbb{R}^n \) of a \( d \) dimensional image \( X \) can be expressed as a mixed \( \ell_{1,2} \) norm

\[
\psi(x) = \| D x \|_{1,2} \equiv \sum_{i=1}^n \| D_i x \|_2, \tag{33}
\]

where \( D_i \) is a \( d \times n \) matrix such that \( D_i x \) is a finite-difference approximation of the gradient at the \( i \)th pixel (\( d = 2 \)) or voxel (\( d = 3 \)), and \( D_i \) is the \( i \)th block-row of
$D \in \mathbb{R}^{dn \times n}$. Note that the definition of $D_i$ depends on both boundary conditions and the finite-difference approximation of the gradient.

The TV seminorm (33) is not everywhere differentiable, but it is easy to compute a subgradient using the chain rule and the following property

$$\partial \|x\|_2 = \begin{cases} \{x/\|x\|_2\} & \|x\|_2 > 0 \\ \{y \mid \|y\|_2 \leq 1\} & \text{otherwise} \end{cases}.$$  \hspace{1cm} (34)

Numerically we may compute a subgradient as

$$\tilde{\nabla} \psi(x) \approx D^T \text{diag}(w_1 I, \ldots, w_n I) - 1 Dx,$$

where $w_i = \max\{\tau, \|D_i x\|_2\}$ or $w_i = \|D_i x\|_2 + \tau$ for some small $\tau > 0$ to avoid dividing by zero or a number close to zero. Note that with the choice $w_i = \max\{\tau, \|D_i x\|_2\}$, the above subgradient approximation can be interpreted as the gradient of a smoothed version of the TV seminorm

$$\psi_\tau(x) = \sum_{i=1}^m \phi_\tau(\|D_i x\|_2),$$

where $\phi_\tau: \mathbb{R} \to \mathbb{R}$ is the scaled Huber penalty

$$\phi_\tau(u) = \begin{cases} (u)^2/(2\tau) & |u| \leq \tau \\ |u| - \tau/2 & \text{otherwise} \end{cases}$$

which is once differentiable.

If we define $g_i$ and $h_i$ as in (32a) and (32c), the resulting incremental methods are similar to the ASD-POCS method of Sidky and Pan [39]. This method seeks a solution to the constrained TV-minimization problem

$$\begin{align*}
\text{minimize} & \quad \|Dx\|_{1,2} \\
\text{subject to} & \quad \|Ax - b\|_2 \leq \gamma \\
x & \in \mathcal{C},
\end{align*}$$

(35)

where $\gamma$ is a constant. The problem (35) is equivalent to the problem (30) (with $\psi(x) = \|Dx\|_{1,2}$) in the sense that for each $\lambda > 0$, there exists a constant $\gamma > 0$ such that both problems have the same set of minimizers. Each iteration of the ASD-POCS method consists of a complete ART cycle and a projection on $\mathcal{C}$, followed by $s$ subgradient steps based on subgradients of the TV seminorm. The method is adaptive in the sense that it adjusts both the step size used in the ART cycle and the step size used for the subgradient steps at each iteration. As a consequence, ASD-POCS does not necessarily converge to a solution to the problem (35), but in practice it often produces a feasible $x$ with low TV quite fast.

As mentioned in the beginning of this section, we obtain another pair of incremental methods if we define the functions $g_i(x)$ and $h_i(x)$ as in (32a) and (32b). The resulting methods do not require a subgradient of the TV seminorm, but instead we need to evaluate the proximal operators $\text{prox}_{t_k g_{m+i}}(x), i = 1, \ldots, s$. If we choose $s = 1$ and $g_{m+1} = \lambda\|Dx\|_{1,2}$, this amounts to solving an unconstrained TV denoising problem, i.e.,

$$\text{prox}_{t_k g_{m+1}}(x) = \arg \min_{u \in \mathbb{R}^n} \left\{ t_k \lambda \|Du\|_{1,2} + 1/2 \|u - x\|_2^2 \right\}. \hspace{1cm} \square$$

\text{Springer}
This is a strongly convex optimization problem that can be solved efficiently with e.g.
FISTA \cite{2} or NESTA \cite{3}. The resulting relaxed incremental proximal algorithms
resemble ART in that every cycle involves damped projections onto the \( m \) hyper-
planes defined by the rows of \( A \), and in addition, every cycle also includes a denoising
step.

5.2 Scaled least-squares with regularization

As a second example, we consider the scaled regularized least-squares problem

\[
\begin{align*}
\text{minimize} & \quad \frac{1}{2} \| M(Atu - b) \|^2_2 + \lambda \psi(Tu) \\
\text{subject to} & \quad Tu \in C
\end{align*}
\]

(36)

with variable \( u \in \mathbb{R}^n \) and where \( M \in \mathbb{R}^{m \times m} \) and \( T \in \mathbb{R}^{n \times n} \) are nonsingular. This
problem is directly related to the SIRT algebraic iterative methods; see \cite{19} for dif-
f erent choices of \( M \) and \( T \). The problem (30) is equivalent to (36) when \( M = I \), and
this follows by making a change of variables \( x = Tu \) in (36). Thus, we may view \( T \) as a preconditioner. In practice the matrix \( T \) should be chosen such that projections
on the set \( C_T = \{ u \mid Tu \in C \} \) are easy to compute, i.e.,

\[
\mathcal{P}_{C_T}(\tilde{u}) = \arg \min_{u \in C} \| u - \tilde{u} \|^2_2 = T^{-1} \arg \min_{v \in C} \| T^{-1} (v - T\tilde{u}) \|^2_2
\]

should be cheap to evaluate. This is the case if, for example, \( C \) is a “box” of the form
\( \{ x \mid c \leq x \leq d \} \) with \( c, d \in \mathbb{R}^n \) and \( T \) is diagonal and positive; in this case we have
\( \mathcal{P}_{C_T}(\tilde{u}) = T^{-1} \mathcal{P}_C(T\tilde{u}) \).

The problem (36) can be used to derive “preconditioned” variants of R-IPG1 and
R-IPG2 for the problem (30). To demonstrate this, we let \( M = I \) and define \( g_i(u) = \frac{1}{2} (a_i^T Tu - b_i)^2 \) and \( h_i(u) = \frac{\lambda}{m} \psi(Tu) \) for \( i = 1, \ldots, m \). The R-IPG1 updates for
the scaled problem (36) can then be expressed as

\[
\begin{align*}
w_k &= u_k - \frac{T^T a_{ik} (a_{ik}^T Tu_k - b_{ik})}{\| Ta_{ik} \|^2_2 + t_k^{-1}} \\
z_k &= w_k - \frac{t_k \lambda}{m} T^T \nabla \psi(Tw_k) \\
u_{k+1} &= \mathcal{P}_{C_T}(\rho z_k + (1 - \rho) u_k)
\end{align*}
\]

(37a)

and after a change of variables \( (x_k = Tu_k, \tilde{w}_k = Tw_k \) and \( \tilde{z}_k = Tz_k \), we obtain the
following “preconditioned” R-IPG1 method for the problem (30)

\[
\begin{align*}
\tilde{w}_k &= x_k - \frac{T T^T a_{ik} (a_{ik}^T x_k - b_{ik})}{\| Ta_{ik} \|^2_2 + t_k^{-1}} \\
\tilde{z}_k &= \tilde{w}_k - \frac{t_k \lambda}{m} T T^T \nabla \psi(\tilde{w}_k) \\
x_{k+1} &= \mathcal{P}_{C_T}(\rho T^{-1} \tilde{z}_k + (1 - \rho) T^{-1} x_k)
\end{align*}
\]

(37b)

(37c)

Although this method converges to a minimizer of (30) for any nonsingular \( T \) (pro-
vided that a diminishing step-size sequence is used), the initial rate of convergence
may vary in practice. Moreover, if $T$ is diagonal, the projection (37c) can be expressed as

$$x_{k+1} = T^{-1}P_C (\rho z_k + (1 - \rho) x_k).$$

(38)

A reasonable preconditioning strategy may be to define $T$ such that all the columns of $AT$ have unit $\ell_p$-norm.

6 Numerical results

The numerical experiments described in this section were carried out in MATLAB, and we used the package AIR Tools [24] to generate the $N \times N$ Shepp-Logan represented by $x^{\text{exact}}$ and the sparse matrix $A$. The underlying model is a parallel-beam tomography problem with $p$ projections, each involving $r$ rays; hence $A$ is $pr \times N^2$.

To avoid what is sometimes referred to as an “inverse crime” [32], we generate the measurement vector $b$ as follows. First, using a larger number of rays $\tilde{r}$ and a finer grid with $\tilde{N} \times \tilde{N}$ pixels, we generate a noise-free sinogram $B \in \mathbb{R}^{\tilde{r} \times p}$; in the experiments we use $\tilde{N} = \text{round}(\sqrt{3}N)$ and $\tilde{r} = \text{round}(\sqrt{2}r)$. We then use interpolation to compute a noise-free sinogram $B \in \mathbb{R}^{r \times p}$ that consists of $r$ samples for each projection angle, and finally we obtain the noisy measurement data $b$ as

$$b = b^{\text{exact}} + e, \quad b^{\text{exact}} = \text{vec}(B).$$

Here $e$ is a normally distributed noise vector with elements $e_i \sim \mathcal{N}(0, \sigma^2)$ and $\sigma$ is chosen such that $\|e\|_2 / \|b^{\text{exact}}\|_2 = \eta$, where we specify the noise level $\eta$.

6.1 The advantage of damped ART

Our first example illustrates the use of the damped ART method (22) from Section 4.2. When the noise level $\eta$ is high, the reconstructed image tends to have large errors in the corners pixels. These pixels correspond to rows of $A$ that have small norm $\|a_i\|_2$, and once such an error has occurred it stays during the following iterations.

A simple and adaptive way to suppress these errors is to use damped ART where we set the parameter $t_k$ in (22) to a constant value chosen such that only updates for rows with small norm are affected. Figure 2 illustrates this for an example for the unrelaxed method ($\rho = 1$) with $\eta = 0.08$, $N = 32$, $r = 32$, and $p = 36$. The top row shows cycles $\ell = 2, 4, 6, 8$ for the standard ART method, and the bottom row shows the same iterates for damped ART with the choice $t_k^{-1} = 0.1 \max_i \|a_i\|^2_2$. The damping clearly suppresses the noise in the corners of the image without affecting the central part, and the results are not sensitive to the factor (here, chosen as 0.1).

6.2 Damped ART with relaxation

In the next experiment, we investigate the role of the relaxation parameter $\rho$ and the parameter $t_k$ for the damped ART method (22). We use a larger test problem with $N = 256$, $r = 362$, $p = 120$, and $\eta = 0.02$. With this geometry, the average row norm $\|a_i\|_2$ is of the order 10.
Fig. 2: Four iterations of standard and damped ART methods for an example with noisy data; the damping $t_k^{-1} = 0.1 \max_i \|a_i\|_2^2$ suppresses the noise in the corners of the image without affecting the central part.

Each of the plots in Fig. 3 shows the norm of the relative error for different values of $\rho$ and with a fixed $t_k$. Observe that when $t_k$ is small, the best performance is achieved with overrelaxation (i.e., $1 < \rho < 2$) whereas when $t_k$ is large, underrelaxation ($0 < \rho < 1$) yields the best result. Note also that with a large $t_k$, the relaxation parameter has a strong influence on best iterate in terms of the minimum error. In particular, the unrelaxed method ($\rho = 1$) is poor for both $t_k = 0.1$ and $t_k = 1.0$, but for $t_k = 0.001$ its performance is similar to that of the overrelaxed methods. Finally, recall that the damped ART method is equivalent to ART if we let $t_k \to \infty$, and in

Fig. 3: The relative error $\|x_{\ell m} - x^{\text{exact}}\|_2/\|x^{\text{exact}}\|_2$ for damped ART with relaxation, after $\ell = 0, 1, 2, \ldots, 10$ full cycles, with different values of the relaxation parameter $\rho$ and a fixed parameter $t_k$. 

$\rho = 0.1$, $\rho = 0.5$, $\rho = 1.0$, $\rho = 1.5$, $\rho = 1.9$.
this example, the damped ART method is practically indistinguishable from ART for $t_k \geq 1.0$.

### 6.3 Generalized ART with regularization

In our last experiment, we consider the TV-regularized reconstruction problem (35) with nonnegativity constraints, i.e., $C = \{x \mid x \geq 0\}$. We use the problem parameters $N = 512, r = 724, p = 60$, and $\eta = 0.01$, and we split $A$ into $p$ blocks $A_1, \ldots, A_p$ (one for each parallel projection) and define

$$g_i(x) = \frac{1}{2} \| A_i x - b_i \|_2^2, \quad h_i(x) = \lambda/p \psi(x), \quad i = 1, \ldots, p.$$  

Here $\psi(x)$ is the TV seminorm defined in (33), and $\lambda > 0$ is the regularization parameter. The proximal operator associated with $g_i(x)$ is given in (24). In this example, the matrix $A_i A_i^T + t_k^{-1} I$ is tridiagonal because of the parallel-beam geometry, and hence the proximal operator can be evaluated efficiently.

To establish a ground truth for the purpose of evaluating the quality of the reconstructions, we first solve the TV-regularized least-squares problem for a number of different regularization parameters using the primal–dual first-order method of Chambolle and Pock [13]. We obtain the best result with $\lambda^* \approx 18$ for which the norm of the relative error is approximately 0.14. We then solve the TV-regularized

![Graph showing relative error for different parameters](image-url)

**Fig. 4** The relative error \( \| x_{\text{ref}} - x_{\text{exact}} \|_2 / \| x_{\text{exact}} \|_2 \) for damped block ART with TV regularization after \( \ell = 0, 1, 2, \ldots, 20 \) full cycles with different values of $\rho$ and $t_0$. The dashed line marks the norm of the relative error of the reference-solution to the TV-regularized least-squares problem obtained using the primal–dual first-order method of Chambolle and Pock [13].
problem with R-IPG1 for different values of $\rho$, using the regularization parameter $\lambda^*$ and cyclic control. Furthermore, we use a diminishing step-size sequence in which the parameter $t_k$ is fixed throughout a complete cycle, i.e., $t_k = t_0 \left\lceil k/p \right\rceil - 1$ for $k = 1, 2, \ldots$, where $t_0$ is the initial value.

The plots in Fig. 4 show our results. We see that the initial rate of convergence strongly depends on the choice of both the relaxation parameter $\rho$ and the initial parameter $t_0$. Nevertheless, with suitable $\rho$ and $t_0$, the method exhibits very fast initial convergence and achieves a reasonably accurate approximate solution after only about 10 cycles. In this example, we obtained the best results with overrelaxation. This can be seen from the plot in Fig. 5 which shows the norm of the relative error after 20 cycles for different values of $\rho$ and $t_0$. It is also clear from this plot that, in this example, the overrelaxed method finds a reasonably accurate approximate solution within 20 cycles when $t_0$ is between approximately 0.01 and 0.1.

7 Conclusions

This work contributes to existing knowledge on row-action methods by providing an extension of the incremental proximal gradient framework of Bertsekas. By adding a relaxation parameter this framework, we have shown that it is possible to interpret many well-known row-action methods for tomographic imaging as incremental methods for solving some convex optimization problem. More importantly, the framework also allows us to derive new generalized row-action methods that are based on generalized projections (i.e., proximal operators). We demonstrated this with several examples, including new ART-like methods for robust regression and regularized regression.

Our numerical experiments suggest that with suitably chosen parameters, the relaxed incremental proximal gradient methods can obtain good approximate solutions in a small number of cycles, even for problems that involve a nonsmooth regularization term such as TV. Interestingly, in most cases we obtained the best
results using either under- or overrelaxation which underlines the practical importance of relaxation. However, the question of how to choose algorithm parameters for a given problem remains a difficult one, and further work is needed to investigate this issue.

Appendix: A Proof of Proposition 1

We start by proving (19) for the iteration (12). Let $y$ denote a vector that belongs to $C$. Then, using the nonexpansiveness of the projection operator $P_C$ and (12c), we have

$$
\|x_{k+1} - y\|^2 \leq \rho^2 \|z_k - y\|^2 + (1 - \rho)^2 \|x_k - y\|^2 + 2\rho(1 - \rho)(z_k - y)^T(x_k - y). 
$$

(39)

It follows from (12a) that $w_k = x_k - t_k \nabla g_{i_k}(w_k)$ for some $\nabla g_{i_k}(w_k) \in \partial g_{i_k}(w_k)$, and combining this with (12b), we get

$$
z_k - y = x_k - y - t_k \nabla g_{i_k}(w_k) - t_k \nabla h_{i_k}(w_k).
$$

Taking inner products with first $z_k - y$ and then $x_k - y$ on both sides of this equation, we obtain the following expression

$$
2(z_k - y)^T(x_k - y) = \|x_k - y\|^2 + \|z_k - y\|^2 + t_k \nabla f_{i_k}(w_k)^T(z_k - x_k)
$$

where $\nabla f_{i_k}(w_k) = \nabla g_{i_k}(w_k) + \nabla h_{i_k}(w_k)$. Using this result in (39) and by substituting $x_k - t_k \nabla f_{i_k}(w_k)$ for $z_k$, we obtain the inequality

$$
\|x_{k+1} - y\|^2 \leq \rho \|x_k - y - t_k \nabla f_{i_k}(w_k)\|^2 + (1 - \rho)\|x_k - y\|^2 + \rho(1 - \rho)t_k^2 \|\nabla f_{i_k}(w_k)\|^2. 
$$

(40)

Expanding the first term on the right-hand side of this inequality yields

$$
\rho \|x_k - y - t_k \nabla f_{i_k}(w_k)\|^2 = \rho \|x_k - y\|^2 + \rho \|\nabla f_{i_k}(w_k)\|^2 - 2\rho t_k \nabla f_{i_k}(w_k)^T(x_k - y)
$$

(41)

and furthermore, using the definition of a subgradient of $f_{i_k}$ at $w_k$, the last term on the right-hand side of (41) can be bounded from above as

$$
-2\rho t_k \nabla f_{i_k}(w_k)^T(x_k - y) \leq -2\rho t_k(f_{i_k}(w_k) - f_{i_k}(y))
+ 2\rho t_k \nabla f_{i_k}(w_k)^T(w_k - x_k). 
$$

(42)

Combining (41), (42), and (40), we get

$$
\|x_{k+1} - y\|^2 \leq \|x_k - y\|^2 - 2\rho t_k(f_{i_k}(w_k) - f_{i_k}(y))
+ \rho^2 t_k^2 \|\nabla f_{i_k}(w_k)\|^2 + 2\rho t_k \nabla f_{i_k}(w_k)^T(w_k - x_k)
$$

(43)

and using (12a) and the definition of $\nabla f_{i_k}(w_k)$, we can express the last two terms on the right-hand side as

$$
\rho t_k^2 \left(\rho \|\nabla h_{i_k}(w_k)\|^2 - (2 - \rho)\|\nabla g_{i_k}(w_k)\|^2 + 2(1 - \rho)\nabla g_{i_k}(w_k)^T \nabla h_{i_k}(w_k)\right).
$$
Using the Cauchy-Schwartz inequality and the inequality $\|\nabla_h i_k(w_k)\|_2 \leq c$ from Assumption 1, we obtain the bound
\[
\rho^2 t_k^2 \|\nabla f_i_k(w_k)\|_2^2 + 2\rho t_k \nabla f_i_k(w_k)^T (w_k - x_k) \\
\leq \rho t_k^2 \left( \rho c^2 - (2 - \rho) \|\nabla g_i_k(w_k)\|_2^2 + 2c|1 - \rho| \|\nabla g_i_k(w_k)\|_2 \right).
\]
This is a concave function of $\|\nabla g_i_k(w_k)\|_2$ for $\rho \in [\delta, 2 - \delta]$, and we obtain a simpler bound by maximizing over $\|\nabla g_i_k(w_k)\|_2 \leq c$. Thus, if we let $\|\nabla g_i_k(w_k)\|_2 = c \min\{|1 - \rho|/(2 - \rho), 1\}$ we obtain the bound
\[
\rho^2 t_k^2 \|\nabla f_i_k(w_k)\|_2^2 + 2\rho t_k \nabla f_i_k(w_k)^T (w_k - x_k) \leq \rho t_k^2 c^2 \alpha(\rho) \tag{44}
\]
where
\[
\alpha(\rho) = \begin{cases} 
1/(2 - \rho) & \delta \leq \rho \leq 3/2 \\
4(\rho - 1) & 3/2 \leq \rho \leq 2 - \delta.
\end{cases}
\]
Combining (43) and (44) yields
\[
\|x_{k+1} - y\|_2^2 \leq \|x_k - y\|_2^2 - 2\rho t_k (f_i_k(w_k) - f_i_k(y)) + \rho t_k^2 c^2 \alpha(\rho). \tag{45}
\]
Applying this bound recursively, and since the index sequence $\{i_k\}$ is cyclic, we have
\[
\|x_{k+m} - y\|_2^2 \leq \|x_k - y\|_2^2 - 2\rho t_k (f(x_k) - f(y)) + m\rho t_k^2 c^2 \alpha(\rho)
\]
\[
+ 2\rho t_k \sum_{j=1}^{m} (f_j(x_k) - f_j(w_{k+j-1})). \tag{46}
\]
We can upper bound $f_j(x_k) - f_j(w_{k+j-1})$ using Assumption 1, i.e.,
\[
f_j(x_k) - f_j(w_{k+j-1}) \leq 2c \|x_k - w_{k+j-1}\|_2. \tag{47}
\]
Furthermore, from the triangle inequality we have
\[
\|x_k - w_{k+j-1}\|_2 \leq \|x_k - x_{k+1}\|_2 + \ldots + \|x_{k+j-2} - x_{k+j-1}\|_2 + \|x_{k+j-1} - w_{k+j-1}\|_2 \tag{48}
\]
and using (39), the first $j - 1$ right-hand side terms can be bounded using the inequality
\[
\|x_k - x_{k+1}\|_2 \leq \rho \|x_k - z_k\|_2 = \rho t_k \|\nabla g_i_k(w_k) + \nabla h_i_k(w_k)\|_2 \leq 2\rho t_k c. \tag{49}
\]
Similarly, from (12a), we have $\|w_k - x_k\|_2 \leq t_k c$, and hence
\[
\|x_k - w_{k+j-1}\|_2 \leq 2(j - 1)\rho t_k c + t_k c \tag{50}
\]
and
\[
2\rho t_k \sum_{j=1}^{m} (f_j(x_k) - f_j(w_{k+j-1})) \leq 4\rho t_k^2 c^2 \sum_{j=1}^{m} (2(j - 1)\rho + 1) = 4\rho t_k^2 c^2 (\rho m^2 + (1 - \rho)m). \tag{51}
\]
Combining (45), (50), and (51), we obtain the desired result (19).
We now prove (19) for the iteration (13). Using (13a) and (13b), we have that
\[ z_k - y = x_k - y - t_k \nabla g_{i_k}(z_k) - t_k \nabla h_{i_k}(x_k). \]
Taking inner products on both sides of this equation with first \( z_k - y \) and then \( x_k - y \), we obtain the equations
\[
(z_k - y)^T (x_k - y) = \|z_k - y\|^2 + t_k (z_k - y)^T (\nabla g_{i_k}(z_k) + \nabla h_{i_k}(x_k))
\]
\[
(z_k - y)^T (x_k - y) = \|x_k - y\|^2 - t_k (x_k - y)^T (\nabla g_{i_k}(z_k) + \nabla h_{i_k}(x_k))
\]
and adding these yields
\[
2(z_k - y)^T (x_k - y) = \|x_k - y\|^2 + \|z_k - y\|^2 + t_k (z_k - x_k)^T (\nabla g_{i_k}(z_k) + \nabla h_{i_k}(x_k)).
\]
Using this in (39) (which holds for both R-IPG1 and R-IPG2), we have that
\[
\|x_{k+1} - y\|^2 \leq \rho \|z_k - y\|^2 + (1 - \rho) \|x_{k+1} - y\|^2 - \rho (1 - \rho) \|z_k - x_k\|^2
\]
where the first term on the right-hand side can be expressed as
\[
\rho \left( \|x_k - y\|^2 + t_k^2 \|\nabla g_{i_k}(z_k) + \nabla h_{i_k}(x_k)\|^2 - 2 (\nabla g_{i_k}(z_k) + \nabla h_{i_k}(x_k))^T (x_k - y) \right)
\]
and using the definition of the subdifferentials \( \partial g_{i_k}(z_k) \) and \( \partial h_{i_k}(x_k) \) together with (44), we obtain
\[
\|x_{k+1} - y\|^2 \leq \|x_k - y\|^2 - 2 \rho t_k (f_{i_k}(x_k) - f_{i_k}(y)) + \rho t_k^2 c^2 \alpha(\rho)
+ 2 \rho t_k (g_{i_k}(x_k) - g_{i_k}(z_k)).
\]
Consequently, after a complete cycle (i.e., \( m \) iterations), we have
\[
\|x_{k+m} - y\|^2 \leq \|x_k - y\|^2 - 2 \rho t_k (f_{i_k}(x_k) - f_{i_k}(y)) + m \rho t_k^2 c^2 \alpha(\rho)
+ 2 \rho t_k \sum_{j=1}^m (f_j(x_k) - f_j(x_{k+j-1}))
+ 2 \rho t_k \sum_{j=1}^m (g_j(x_{k+j-1}) - g_j(z_{k+j-1})).
\]
Now, using Assumption 2, we can bound \( f_j(x_k) - f_j(x_{k+j-1}) \) above as
\[
f_j(x_k) - f_j(x_{k+j-1}) \leq 2c \|x_k - x_{k+j-1}\|_2
\]
where
\[
\|x_k - x_{k+j-1}\|_2 \leq \|x_k - x_{k+1}\|_2 + \cdots + \|x_{k+j-2} - x_{k+j-1}\|_2
\]
and
\[
\|x_k - x_{k+1}\|_2 \leq \rho \|z_k - x_k\|_2 = \rho t_k \|\nabla g_{i_k}(z_k) + \nabla h_{i_k}(x_k)\|_2 \leq 2 \rho t_k c
\]
and this implies that
\[
2 \rho t \sum_{j=1}^{m} \left( f_j(x_k) - f_j(x_{k+j-1}) \right) \leq 4 \rho^2 t_k^2 c^2 (m^2 - m). \tag{56}
\]

Similarly, using Assumption 2, we have that
\[
g_{ik}(x_{k+j-1}) - g_{ik}(z_{k+j-1}) \leq c \left\| x_{k+j-1} - z_{k+j-1} \right\|_2 \\
\leq t_k c \left\| \nabla g_{ik}(z_{k+j-1}) + \nabla h_{ik}(x_{k+j-1}) \right\|_2 \\
\leq 2t_k c^2
\]
and hence
\[
2 \rho t \sum_{j=1}^{m} (g_j(x_{k+j-1}) - g_j(z_{k+j-1})) \leq 4m \rho t_k^2 c^2. \tag{57}
\]

Combining (55), (56), and (57), we get the desired result (19).

References

1. Albert, A.: Regression and the Moore-Penrose Pseudo Inverse. Academic, New York (1972)
2. Beck, A., Teboulle, M.: A fast iterative shrinkage-thresholding algorithm for linear inverse problems. SIAM J. Imaging Sci. 2(1), 183–202 (2009)
3. Becker, S., Bobin, J., Candès, E.: NESTA: A fast and accurate first-order method for sparse recovery. SIAM J. Imaging Sci. 4(1), 1–39 (2011)
4. Bertero, M., Boccacci, P.: Introduction to Inverse Problems in Imaging. Institute of Physics Publishing, Bristol (1998)
5. Bertsekas, D.P.: A new class of incremental gradient methods for least squares problems. SIAM J. Optim. 7(4), 913–926 (1997)
6. Bertsekas, D.P.: Nonlinear Programming, 2nd edn. Athena Scientific, Nashua (1999)
7. Bertsekas, D.P.: Incremental gradient, subgradient, and proximal methods for convex optimization: A survey. In: Sra, S., Nowozin, S., Wright, S.J. (eds.) Optimization for Machine Learning, pp. 85–119. MIT Press, Cambridge, MA (2011)
8. Bertsekas, D.P.: Incremental proximal methods for large scale convex optimization. Math. Program. 129, 163–195 (2011)
9. Blatt, D., Hero, A., Gauchman, H.: A convergent incremental gradient method with a constant step size. SIAM J. Optim. 18(1), 29–51 (2007)
10. Censor, Y., Davidi, R., Herman, G.T.: Perturbation resilience and superiorization of iterative algorithms. Inverse Probl. 26(6), 065,008 (2010)
11. Censor, Y., Eggermont, P., Gordon, D.: Strong underrelaxation in Kaczmarz’s method for inconsistent systems. Numer. Math. 41(1), 83–92 (1983)
12. Chambolle, A., De Vore, R.A., Lee, N.Y., Lucier, B.J.: Nonlinear wavelet image processing: variational problems, compression, and noise removal through wavelet shrinkage. IEEE Trans. Image Process. 7(3), 319–335 (1998)
13. Chambolle, A., Pock, T.: A first-order primal-dual algorithms for convex problems with applications to imaging. J. Math. Imaging Vis. 40, 120–145 (2011)
14. Combettes, P.L., Pesquet, J.C.: Proximal splitting methods in signal processing. In: Bauschke, H., Burachik, R., Combettes, P., Elser, V., Luke, D. (eds.) Fixed-Point Algorithms for Inverse Problems in Science and Engineering, pp. 185–212. Springer, New York, NY (2011)
15. Daubechies, I., Defrise, M., De Mol, C.: An iterative thresholding algorithm for linear inverse problems with a sparsity constraint. Commun. Pure Appl. Math. 57(11), 1413–1457 (2004)
16. Davidi, R., Herman, G.T., Censor, Y.: Perturbation-resilient block-iterative projection methods with application to image reconstruction from projections. Int. Trans. Oper. Res. 16(4), 505–524 (2009)
17. Dielman, T.E.: Least absolute value regression: Recent contributions. J. Stat. Comput. Simul. **75**(4), 263–286 (2005)
18. Elfving, T.: Block-iterative methods for consistent and inconsistent linear equations. Numer. Math. **35**, 1–12 (1980)
19. Elfving, T., Hansen, P.C., Nikazad, T.: Semiconvergence and relaxation parameters for projected sirt algorithms. SIAM J. Sci. Comput. **34**(4), A2000–A2017 (2012)
20. Elfving, T., Hansen, P.C., Nikazad, T.: Semi-convergence properties of Kaczmarz’s method. Submitted to Inverse Problems (2013)
21. Elfving, T., Nikazad, T., Popa, C.: A class of iterative methods: Semi-convergence, stopping rules, inconsistency, and constraining. In: Censor, Y., Jiang, M., Wang, G. (eds.) Biomedical Mathematics: Promising Directions in Imaging, Therapy Planning, and Inverse Problems. Medical Physics Publishing, Madison (2010)
22. Friedlander, M., Schmidt, M.: Hybrid deterministic-stochastic methods for data fitting. SIAM J. Sci. Comput. **34**(3), A1380–A1405 (2012)
23. Gordon, R., Bender, R., Herman, G.T.: Algebraic reconstruction techniques (ART) for three-dimensional electron microscopy and X-ray photography. J. Theor. Biol. **29**(3), 471–481 (1970)
24. Hansen, P.C., Saxild-Hansen, M.: AIR Tools—a MATLAB package of algebraic iterative reconstruction methods. J. Comput. Appl. Math. **236**(8), 2167–2178 (2012)
25. Herman, G.T.: Fundamentals of Computerized Tomography: Image Reconstruction from Projections, 2nd edn. Springer, New York (2009)
26. Herman, G.T., Meyer, L.B.: Algebraic reconstruction techniques can be made computationally efficient. IEEE Trans. Med. Imaging **12**, 600–609 (1993)
27. Jiang, M., Wang, G.: Convergence studies on iterative algorithms for image reconstruction. IEEE Trans. Med. Imaging **22**(5), 569–579 (2003)
28. Kaczmarz, S.: Angenäherte auflösung von systemen linearer gleichungen. Bulletin International de l’Académie Polonaise des Sciences et des Lettres **35**, 355–357 (1937)
29. Martinet, B.: Régularisation d’inéquations variationnelles par approximations successives. Revue Française d’Informatique et de Recherche Opérationnelle **4**(3), 154–158 (1970)
30. Martinet, B.: Algorithmes pour la resolution de problems d’optimisation et de minimax. Ph.D. thesis (1972)
31. Moreau, J.J.: Proximité et dualité dans un espace hilbertien. Bull. Math. Soc. France **93**, 273–299 (1965)
32. Mueller, J., Siltanen, S.: Linear and Nonlinear Inverse Problems with Practical Applications. SIAM, Philadelphia, PA (2012)
33. Natterer, F.: The Mathematics of Computerized Tomography. SIAM, Philadelphia (2001)
34. Nesterov, Y.: Introductory Lectures on Convex Optimization. Kluwer Academic Publishers, Dordrecht (2004)
35. Recht, B., Re, C.: Toward a noncommutative arithmetic-geometric mean inequality: conjectures, case-studies, and consequences. In: Proceedings of the 25th Annual Conference on Learning Theory (2012)
36. Rockafellar, R.T.: Convex Analysis, 2nd edn. Princeton University Press, Princeton (1970)
37. Rockafellar, R.T.: Monotone operators and the proximal point algorithm. SIAM J. Control Optim. **14**(5), 877–898 (1976)
38. Rudin, L., Osher, S.J., Fatemi, E.: Nonlinear total variation based noise removal algorithms. Phys. D **60**, 259–268 (1992)
39. Sidky, E.Y., Pan, X.: Image reconstruction in circular cone-beam computed tomography by constrained, total-variation minimization. Phys. Med. Biol. **53**(17), 4777 (2008)