Inexact variable metric proximal gradient methods with line-search for convex and nonconvex optimization

Silvia Bonettini

Dipartimento di Scienze Fisiche, Informatiche e Matematiche
Università di Modena e Reggio Emilia

Optimization Algorithms and Software for Inverse problems
www.oasis.unimore.it

Computational Methods for Inverse Problems in Imaging
Como, 16-18 July 2018
Collaborators and main references

Joint works with:

- Marco Prato, Università di Modena e Reggio Emilia
- Federica Porta, Simone Rebegoldi, Valeria Ruggiero, Università di Ferrara
- Ignace Loris, Université Libre de Bruxelles

Main references:

Several optimization problems arising from the Bayesian approach to inverse problems have the following structure

$$
\min_{x \in \mathbb{R}^n} f(x) \equiv f_0(x) + f_1(x),
$$

where:

- $f_0(x)$ continuously differentiable, possibly nonconvex.
  - usually expressing some kind of data discrepancy
- $f_1(x)$ convex, possibly nondifferentiable
  - usually expressing regularization

Goal: develop a numerical optimization algorithm producing a good approximation of the solution of the minimization problem in few, cheap iterations.
The class of proximal gradient methods

Proximal gradient methods, aka forward-backward methods, exploit the smoothness of $f_0$ and the convexity of $f_1$ in problem

$$\min_{x \in \mathbb{R}^n} f(x) \equiv f_0(x) + f_1(x),$$

**Definition (Proximal gradient method)**

Any first order method based on the following two operations:

- **Explicit Forward/Gradient step**: computation of the gradient $\nabla f_0(x)$
- **Implicit Backward/Proximal step**: computation of the *proximity* (or resolvent) operator:

$$\text{prox}_{f_1}(z) = \arg \min_{x \in \mathbb{R}^n} f_1(x) + \frac{1}{2} \|x - z\|^2$$

**Example**: If $\Omega \subset \mathbb{R}^n$ is a closed convex set, we can define the indicator function

$$\iota_\Omega(x) = \begin{cases} 0 & \text{if } x \in \Omega \\ +\infty & \text{otherwise} \end{cases} \Rightarrow \text{prox}_{\iota_\Omega}(z) = \Pi_\Omega(z)$$

(orthogonal projection onto $\Omega$).

**NB**: gradient projection methods are special instances of proximal gradient methods.

Silvia Bonettini

Inexact variable metric proximal gradient methods
A basic forward-backward scheme

\[ z^{(k)} = x^{(k)} - \alpha_k \nabla f_0(x^{(k)}) \quad \leftarrow \quad \text{Forward step} \]
\[ y^{(k)} = \text{prox}_{\alpha_k f_1}(z^{(k)}) \quad \leftarrow \quad \text{Backward step} \]
\[ d^{(k)} = y^{(k)} - x^{(k)} \]
\[ x^{(k+1)} = x^{(k)} + \lambda_k d^{(k)} \]

**NB:** The steplength parameters \( \alpha_k, \lambda_k \in \mathbb{R}_{>0} \), in standard convergence analysis, are related to the Lipschitz constant \( L \) of \( \nabla f_0(x) \) [Combettes-Wajs 2006], [Combettes, Wu, 2014] requiring that

\[ \alpha_k \text{ and/or } \lambda_k \leq \frac{C}{L} \]

A motivating problem: nonnegative image restoration from Poisson data

\[
\min_{x \in \mathbb{R}^n} KL(H x, g) + \rho \| \nabla x \| + \iota_{\mathbb{R}^n_{\geq 0}}(x) \quad \text{where} \quad KL(t, g) = \sum_{i=1}^{n} \log \left( \frac{g_i}{t_i} \right) + t_i - g_i
\]

- either \( \nabla f_0 \) is not Lipschitz or \( L \) is very large
- \( \text{prox}_{f_1} \) is not available in closed form
We propose to compute $\lambda_k$ with a line–search approach, starting from 1 and backtracking until a sufficient decrease of the objective function is obtained.

**Generalized Armijo rule [Tseng, Yun, 2009, Porta, Loris, 2015, B. et al., 2016]**

$$f(x^{(k)} + \lambda_k d^{(k)}) \leq f(x^{(k)}) + \beta \lambda_k h^{(k)}(y^{(k)}),$$

where $\beta \in (0, 1)$ and

$$h^{(k)}(y) = \nabla f_0(x^{(k)})^T (y - x^{(k)}) + \frac{1}{2\alpha_k} \| y - x^{(k)} \|^2 + f_1(y) - f_1(x^{(k)}) ,$$

**NB1:** We have $y^{(k)} = \text{prox}_{\alpha_k f_1}(x^{(k)} - \alpha_k \nabla f_0(x^{(k)})) = \arg \min_{y \in \mathbb{R}^n} h^{(k)}(y)$. Since $h^{(k)}(y^{(k)}) < 0$, we obtain a monotone decrease of the objective function.

**NB2:** For $f_1 \equiv 0$, dropping the quadratic term we obtain the standard Armijo rule for smooth optimization.

**Pros:**
- No need of any Lipschitz assumption
- Adaptive selection of $\lambda_k$ (no user provided parameter)
- No assumptions on $\alpha_k$, just be bounded above and away from zero.

**Cons:**
- Needs the evaluation of the function $f$ at each backtracking loop (usually 1-2 per outer iteration).
Inexact computation of the proximity operator (1)

Basic idea

Compute an approximation \( \tilde{y}^{(k)} \) of \( y^{(k)} \) by applying an iterative optimization method to the minimum problem defining the proximity operator:

\[
\tilde{y}^{(k)} \approx y^{(k)} = \arg \min_{y \in \mathbb{R}^n} h^{(k)}(y)
\]

with an increasing accuracy as \( k \) increases. This results in a two loop algorithm and the question now is

How to stop the inner iterations to preserve the convergence of the iterates \( \{x^{(k)}\} \) to a solution?

We need to define a criterion to measure the accuracy of the approximate proximity operator computation.

Crucial properties of this criterion:

- It has to preserve the convergence properties of the whole scheme.
- It must be based on computable quantities.

Borrowing the ideas in [Salzo, Villa, 2012], [Villa et al. 2013]

\[
\text{replace } 0 \in \partial h^{(k)}(y^{(k)}) \text{ with } 0 \in \partial_{\epsilon_k} h^{(k)}(\tilde{y}^{(k)})
\]
Assume that $f_1(x) = g(Ax)$, $A \in \mathbb{R}^{m \times n}$ (easy generalization to $f_1(x) = \sum_{i=1}^{P} g_i(A_ix)$).

The dual problem of the proximity operator computation is

$$
\min_{x \in \mathbb{R}^n} h^{(k)}(x) = \max_{v \in \mathbb{R}^m} \Psi^{(k)}(v) \equiv -\frac{1}{2\alpha_k} \|\alpha_k A^T v - z^{(k)}\|^2 - g^*(v) + C_k
$$

where $g^*$ is the Fenchel convex conjugate of $g$.

If $v^{(k)} = \arg \max \Psi^{(k)}(v)$, then $y^{(k)} = z^{(k)} - \alpha_k A^T v^{(k)}$.

Compute $\tilde{y}^{(k)}$ as follows:

- apply a maximization method to the dual problem, generating the dual sequence $\{v^{(k,\ell)}\}_{\ell \in \mathbb{N}}$ converging to $v^{(k)}$
- compute the corresponding primal sequence $\{\tilde{y}^{(k,\ell)}\}_{\ell \in \mathbb{N}}$, with formula $\tilde{y}^{(k,\ell)} = z^{(k)} - \alpha_k A^T v^{(k,\ell)}$
- stop the inner iterations when

$$
\quad h^{(k)}(\tilde{y}^{(k,\ell)}) - \Psi^{(k)}(v^{(k,\ell)}) \leq \epsilon_k
$$

where

$$
\epsilon_k = \begin{cases} 
\frac{C}{k^q} & \text{with } q > 1 \quad \text{prefixed sequence choice} \\
\eta h^{(k)}(\tilde{y}^{(k,\ell)}) & \text{with } \eta \in (0, 1] \quad \text{adaptive choice}
\end{cases}
$$
Add a new parameter, a s.p.d. scaling matrix $D_k$ which determines a different metric at each iterate:

$$\text{replace } \|x\| \text{ with } \|x\|_{D_k} = x^T D_k x$$

### Variable Metric Inexact Line–Search Algorithm (VMILA)

$$z^{(k)} = x^{(k)} - \alpha_k D_k^{-1} \nabla f_0(x^{(k)}) \quad \leftarrow \text{Scaled Forward step}$$

$$\tilde{y}^{(k)} \approx \text{prox}_{\alpha_k f_1}^{D_k}(z^{(k)}) \equiv y^{(k)} \quad \leftarrow \text{Scaled Inexact Backward step}$$

$$d^{(k)} = \tilde{y}^{(k)} - x^{(k)}$$

$$x^{(k+1)} = x^{(k)} + \lambda_k d^{(k)} \quad \leftarrow \text{Armijo-like line–search}$$
Summary of convergence results about VMILA

**VMILA**

$\lambda_k$ with line–search + inexact computation of the proximal point with increasing accuracy + $\alpha_k$ bounded

**Convex case**

Assumption: $D_k \xrightarrow{k \to \infty} I$ like $C/k^p$, $p > 1$

- Convergence to a minimizer (without Lipschitz assumptions on $\nabla f_0(x)$)
- Convergence rate $f(x^{(k)}) - f^* = \mathcal{O}(1/k)$ (proof with Lipschitz assumptions on $\nabla f_0(x)$)

**Nonconvex case**

Assumption: $D_k$ has bounded eigenvalues.

- Every accumulation point of $\{x^{(k)}\}_{k \in \mathbb{Z}}$ is a stationary point
- If $f$ satisfies the Kurdyka–Lojasiewicz property and $\nabla f_0$ is locally Lipschitz, then $\{x^{(k)}\}_{k \in \mathbb{Z}}$ converges to a stationary point (with exact proximal point computation).

Block-coordinate version of VMILA proposed in [B., Prato, Rebegoldi, 2018, to appear].

**NB:** $\alpha_k$ and $D_k$ are required only to be bounded $\Rightarrow$ use them to implement some acceleration strategy.
Metric selection - A Majorization-Minimization approach

No theoretical results (same rate and lower complexity bound than nonscaled methods). No general recipe for selecting \( D_k \).

Freedom to choose \( D_k, \alpha_k \) allows to use them for accelerating practical performances. Good numerical results with problem dependent practical strategies for \( D_k \).

### Majorization-Minimization idea

Define \( D_k \) such that

\[
x^{(k)} - D_k^{-1} \nabla f_0(x^{(k)}) = \arg \min_{x \in \mathbb{R}^n} F(x, x^{(k)})
\]

where \( F(x, x^{(k)}) \) is a (non necessarily quadratic) auxiliary function for \( f_0 \), i.e.

\[
F(x^{(k)}, x^{(k)}) = f_0(x^{(k)}) \quad \text{and} \quad F(x, x^{(k)}) \geq f_0(x) \quad \forall x \in \mathbb{R}^n
\]

\[
F(\bar{x}) \leq F(\bar{x}, x^{(k)}) \leq F(x^{(k)}, x^{(k)}) = f_0(x^{(k)})
\]

This produces a diagonal \( D_k \) whose elements are obtained from the components of \( \nabla f_0(x^{(k)}) \) in several relevant cases (discrepancy functions for Gaussian, Poisson, Cauchy, multiplicative noise,...) [Yang, Oja, 2011], [Chouzenoux, Pesquet, 2016].

The convergence condition \( D_k \to I \) can be fulfilled by squeezing the elements of the diagonal matrix \( D_k \) to 1 as \( k \) increases.
Given $D_k$, we would choose $\alpha_k$ such that

$$\frac{1}{\alpha_k} D_k \simeq \nabla^2 f_0(x^{(k)})$$

simulating the Taylor’s equality

$$\nabla f_0(x + d) = \nabla f_0(x) + \int_0^1 \nabla^2 f_0(x + td) dt$$

$$\nabla f_0(x^{(k)}) - \nabla f_0(x^{(k-1)}) \simeq \frac{1}{\alpha_k} D_k (x^{(k)} - x^{(k-1)})$$

$$\alpha_k^{BB1} = \arg \min_\alpha \| \frac{1}{\alpha} D_k s^{(k-1)} - w^{(k-1)} \| = \frac{\| D_k s^{(k-1)} \|^2}{s^{(k-1)T} D_k w^{(k-1)}}$$

$$\alpha_k^{BB2} = \arg \min_\alpha \| s^{(k-1)} - \alpha D_k^{-1} w^{(k-1)} \| = \frac{s^{(k-1)T} D_k^{-1} w^{(k-1)}}{\| D_k w^{(k-1)} \|^2}$$

- Good results when the two values are alternated following an adaptive switching rule and projected onto a given interval $[\alpha_{\min}, \alpha_{\max}]$, with $0 < \alpha_{\min} < \alpha_{\max}$.
- Recent developments in steplength selection rules: Ritz values [Fletcher 2012].
VMILA has been tested on a variety of convex and nonconvex image restoration problems.

The numerical comparison shows that its performances are comparable with the ones of state-of-the-art methods such as: Chambolle-Pock (CP) method, preconditioned CP, ADMM, PidSplit+, iPiano, VMFB, FISTA...

Illustration of the effects of a well tailored choice of $\alpha_k, D_k$: nonnegative image deconvolution in presence of Poisson noise with smooth TV regularization.

\[ f(x^{(k)}) - f^* \]
Two acceleration techniques: Extrapolation

**FISTA iteration [Beck, Teboulle, 2008]**

\[
\begin{align*}
\bar{x}^{(k)} &= x^{(k)} + \gamma_k (x^{(k)} - x^{(k-1)}) \quad \leftarrow \text{Extrapolation step} \\
z^{(k)} &= \bar{x}^{(k)} - \alpha_k \nabla f_0(\bar{x}^{(k)}) \quad \leftarrow \text{Forward step} \\
x^{(k+1)} &= \text{prox}_{\alpha_k f_1}(z^{(k)}) \quad \leftarrow \text{Backward step}
\end{align*}
\]

- It applies when \(f_0, f_1\) are both convex and \(\nabla f_0\) is \(L\)-Lipschitz continuous.
- \(\alpha_k\) can be chosen as \(\alpha_k = \alpha \leq \frac{1}{L}\) or with a backtracking procedure, starting from \(\alpha_{k-1}\), guaranteeing that

\[
f_0(x^{(k+1)}) \leq f_0(\bar{x}^{(k)}) + \nabla f_0(\bar{x}^{(k)})^T (x^{(k+1)} - \bar{x}^{(k)}) + \frac{1}{2\alpha_k} \|x^{(k+1)} - \bar{x}^{(k)}\|^2
\]

- Properly chosen extrapolation parameter \(\gamma_k \xrightarrow{k \to \infty} 1\)
  \[
  \gamma_k = \frac{k - 1}{k + a}, \quad a \geq 2
  \]

- Quadratic convergence rate \(f(x^{(k)}) - f^* = O(1/k^2)\) with \(a \geq 2\) and \(f(x^{(k)}) - f^* = o(1/k^2)\) with \(a > 2\) [Attouch-Peypouquet, 2016]

- Convergence of the iterates to a minimizer [Chambolle-Dossal, 2015] with \(a > 2\).
Combining Extrapolation with Scaling and Inexact computation of proximity operator
[B., Porta, Ruggiero, 2016], [B., Rebegoldi, Ruggiero, 2018, submitted]

Scaled, inexact FISTA-like method

\[ \bar{x}^{(k)} = \Pi_{\text{dom}(f)}^D (x^{(k)} + \gamma_k (x^{(k)} - x^{(k-1)})) \leftarrow \text{Extrapolation step with scaled projection} \]

\[ z^{(k)} = \bar{x}^{(k)} - \alpha_k D_k^{-1} \nabla f_0 (\bar{x}^{(k)}) \leftarrow \text{Scaled Forward step} \]

\[ x^{(k+1)} \approx \text{prox}^{D_k}_{\alpha_k f_1} (z^{(k)}) \leftarrow \text{Scaled Inexact Backward step} \]

Main features:

FISTA-like extrapolation + projection + line–search computation of \( \alpha_k \) + inexact computation of the proximity operator + \( D_k \xrightarrow{k \to \infty} I \)

Convex case (\( \nabla f_0 (x) \) Lipschitz continuous):

- Quadratic convergence rate \( f(x^{(k)}) - f^* = \mathcal{O}(1/k^2) \) with \( a \geq 2 \) and \( f(x^{(k)}) - f^* = o(1/k^2) \) with \( a > 2 \)
- Convergence of the iterates to a minimizer with \( a > 2 \).
Numerical results

Nonnegative image deconvolution with Total Variation regularization from Poisson data.

Nonscaled (blue) vs. scaled (red, black)

inexact FISTA-like method

Scaled inexact FISTA-like method (blue) vs.

state of the art

\[
(f(x(k)) - f^*) / f^*
\]

Relative error on the objective function
Conclusions and perspectives

- Line–search based algorithms allow the implementation of FB methods without the knowledge of the Lipschitz constant or even to non Lipschitz problems
  - To do: study a less expensive line–search strategy for the inexact FISTA, to avoid the computation of a new approximate proximal point at each trial step
- The approximation of the proximity operator with an inner loop can be done in such a way that the basic convergence properties of the FB algorithms are not affected
  - To do: investigate other implementable criteria, especially for nonconvex problems.
- Variable metric techniques can be considered as acceleration techniques improving the behaviour of FB methods especially at the first iterations
  - To do: give a better insight on the Majorization-Minimization techniques

Codes available on www.oasis.unimore.it