Multilevel FISTA for image restoration
Guillaume Lauga, Elisa Riccietti, Nelly Pustelnik, Paulo Gonçalves

To cite this version:
Guillaume Lauga, Elisa Riccietti, Nelly Pustelnik, Paulo Gonçalves. Multilevel FISTA for image restoration. ICASSP, IEEE, Jun 2023, Rhodes, Greece. 10.48550/arXiv.2210.15940. hal-03831180v3

HAL Id: hal-03831180
https://hal.science/hal-03831180v3
Submitted on 16 Mar 2023

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L’archive ouverte pluridisciplinaire HAL, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d’enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

Distributed under a Creative Commons Attribution 4.0 International License
MULTILEVEL FISTA FOR IMAGE RESTORATION

Guillaume Lauga†, Elisa Riccietti†, Nelly Pustelnik*, Paulo Gonçalves†

†Univ Lyon, Inria, EnsL, UCBL, CNRS, LIP, UMR 5668, F-69342, Lyon Cedex 07, France
*Ens de Lyon, CNRS, Laboratoire de Physique, F-69342, Lyon, France

ABSTRACT
This paper presents a multilevel fast iterative soft thresholding algorithm (FISTA), based on the use of the Moreau envelope to incorporate correction from coarse models, which is easy to compute when the explicit form of the proximal operator for the considered functions is known. This approach is supported by strong theoretical guarantees: we prove both the rate of convergence and the convergence of the iterates to a minimum in the convex case, an important result for ill-posed problems. We evaluate our approach on image restoration problems and we show that it outperforms classical FISTA for large-scale images.

Index Terms—multilevel optimization, inertial methods, image restoration, proximal methods.

1. INTRODUCTION
Many problems in signal and image processing involve minimizing a sum of a data fidelity term $f$ and a regularization function $g$, formally:

$$
\min_{x \in \mathcal{H}} F(x) := f(x) + g(x)
$$

where $\mathcal{H}$ is a real Hilbert space ($\mathcal{H} = \mathbb{R}^N$ in the following), $f : \mathcal{H} \to (-\infty, +\infty]$ and $g : \mathcal{H} \to (-\infty, +\infty]$ belong to $\Gamma_0(\mathcal{H})$ the class of convex, lower semi-continuous, and proper functions. Moreover, $f$ is supposed to be differentiable with gradient $L_f$-Lipschitz and $F$ is supposed to be coercive.

In the context of restoration, we aim to recover a good quality image from a degraded image $z = Ax + \epsilon$, where $A \in \mathbb{R}^{N \times N}$ models a linear degradation operator and $\epsilon$ stands for the additive noise. To solve this ill-posed problem, we generally consider a regularized least squares formulation, where we denote $g$ the regularization function allowing us to choose the properties that we wish to impose on the solution. A usual choice is to apply the $l_1$-norm on the coefficients raised by a linear transformation $W \in \mathbb{R}^{K \times N}$ (wavelets, frames, dictionary, …), thus promoting the sparsity of the solution [2].

Given a regularization parameter $\lambda > 0$, the associated minimization problem reads:

$$
\hat{x} \in \text{Argmin}_{x \in \mathbb{R}^N} \frac{1}{2} \|Ax - z\|^2 + \lambda \|Wx\|_1.
$$

Many algorithms have been proposed in the literature to estimate $\hat{x}$ (cf. [3–6]). They suffer from a significant increase in computational time with the dimension. Preconditioning techniques can be investigated but generally require strong assumptions for the choice of the preconditioning matrix (e.g., diagonal matrix) leading to limited gains. For the solution of large-scale problems with smooth objective function, it is possible to take advantage of the local structure of the optimization problem (cf. [7] or [8]).

In this paper, we focus on a different family of approaches, the multilevel schemes, which exploit different resolutions of the same problem. In such methods the objective function is approximated by a sequence of functions defined on reduced dimensional spaces (coarser spaces). The descent step is thus calculated at coarser levels with minimal cost and then projected to the fine levels.

These approaches have been mainly studied for the solution of partial differential equations (PDEs), in which $f$ and $g$ are supposed to be differentiable [9, 10], but recently this idea has also been exploited in [11–13] to define multilevel forward-backward proximal algorithms applicable to problem (1) in the case where $g$ is non differentiable.

In this paper we propose a variant of these methods, which we call MMFISTA providing a multilevel alternative to inertial strategies such as FISTA [14, 15]. Our framework relies on the Moreau envelope to define smooth coarse approximations of $g$, which can be easily constructed when the proximal operator of $g$ is known in explicit form. Furthermore, we show under mild assumptions that the convergence guarantees of FISTA hold for MMFISTA, and in particular the convergence of the iterates, an important result for ill-posed problems and, to our knowledge, never established for multilevel inertial proximal methods.

The paper is organized as follows. In Section 2, we recall the main principles of FISTA. Then, we describe MMFISTA. In Section 3, we present its convergence guarantees. Finally,
in Section 4, we present numerical results to confirm the good
behaviour of MMFISTA in an image restoration context.

2. MULTILEVEL FISTA

FISTA – Among the numerous algorithms designed to solve a
minimization problem of the form (1), the most standard stra-
gy, we consider the two-level case: we index by
\( I \) (AD) in the following.

**Multilevel framework** – The multilevel framework exploits
a hierarchy of objective functions, which are representations of
\( F \) at different resolutions and alternate minimization be-
tween these objective functions (following a V cycle procedure [9]).
Without loss of generality and for the sake of clarity, we consider the two-level case: we index by \( h \) (resp. \( H \))
all quantities defined at the fine (resp. coarse) level. We thus
define \( F_h : \mathbb{R}^{N_h} \rightarrow (-\infty, +\infty] \) the objective function
at the fine level where \( N_h = N \), involving \( f_h := f \) and \( g_h := g \). Its approximation at the coarse level is de-
noted \( F_H : \mathbb{R}^{N_H} \rightarrow (-\infty, +\infty] \) where \( N_H < N_h \), which
involves \( f_H \) and \( g_H \). We also define transfer information
operators: a linear operator \( I_h^H : \mathbb{R}^{N_h} \rightarrow \mathbb{R}^{N_H} \) that sends
information from the fine level to the coarse level, and conversely \( I_h^H : \mathbb{R}^{N_H} \rightarrow \mathbb{R}^{N_h} \) that sends information from the
course level back to the fine level. It is classical to choose
\( I_h^H = \eta(I_h^H)^T \), with \( \eta > 0 \).

In a multilevel scheme, we improve the intermediate iterate \( y^k_h \) by performing iterations at the coarse level: \( y^k_h \) is pro-
ected to the coarse level with \( I_h^H \) (5a), a sequence \( (x^k_{h,\ell})_{\ell \in \mathbb{N}} \)
is defined (where \( k \) represents the current iteration at the fine
level and \( \ell \) indexes the iterations at the coarse level) such that:
\[ x^k_{h,\ell+1} = \Phi_h^{k,\ell}(x^k_{h,\ell}) \text{, with } \Phi_h^{k,\ell} \text{ any operator such that} \]
\[ F_H(x^k_{h,m}) \leq F_H(x^k_{h,0}) \text{ for some } m > 0 \]. This yields after \( m \)
iterations at the coarse level (5b) to a step being brought back
at the fine level (5c).

Then, the generic iteration \( k \) of a multilevel method reads :

\[
x^k_{h,0} = I_h^H y^k_h
\]

(5a)

\[
x^k_{h,m} = \Phi_h^{k,m-1} \ldots \Phi_h^{k,0}(x^k_{h,0})
\]

(5b)

\[
y^k_h = y^k_h + \tau_h \gamma_h \nabla f_h(y^k_h)
\]

(5c)

\[
x^k_{h+1} = \text{prox}_{\tau_h g_h}(y^k_h - \tau_h \nabla f_h(y^k_h))
\]

(5d)

\[
y^k_{h+1} = x^k_{h+1} + \alpha_h \gamma_h (x^k_{h+1} - x^k_h)
\]

(5e)

By taking \( x^k_{h,m} = x^k_{h,0} \) one recovers the standard FISTA it-
eration. To ensure that the correction term \( x^k_{h,m} - x^k_{h,0} \) once
projected from coarse level to fine level, induces a decrease of
\( F_h \), we need to appropriately choose :

- the coarse model \( F_H \),
- the minimization scheme \( \Phi_H \).

**Coarse model** \( F_H \) – The coarse iterations are built using
the Moreau envelope of \( g_h \) and of its coarse approximation
\( g_H \). The Moreau envelope provides a natural choice to extend
ideas coming from the classical smooth case [10] to proximal
gradient methods because of its smoothness and its expres-
sion involving the proximity operator. We first recall that for
\( \gamma > 0 \) and \( g \) being a convex, lower semi-continuous, and
proper function of \( \mathcal{H} \) in \((-\infty, +\infty]\), its Moreau envelope,
\( \gamma g \), is the convex, continuous, real-valued function defined by

\[
\gamma g = \inf_{y \in \mathcal{H}} g(y) + (1/2\gamma)\| \cdot - y \|^2,
\]

(6)

which can be expressed explicitly with \( \text{prox}_{\gamma g} \) [16, Remark
12.24]. The gradient of \( \gamma g \) is \( \gamma^{-1} \)-lipschitz and such that
(Prop. 12.30 in [16])

\[
\nabla(\gamma g) = \gamma^{-1}(\text{Id} - \text{prox}_{\gamma g}).
\]

(7)

At iteration \( k \), the coarse model \( F_H \) is defined as

\[
F_H(x^k_h) = f_h(x^k_h) + g_h(x^k_h) + \langle v_{H,k}, x^k_h \rangle
\]

(8)

where

\[
v_{H,k} = I_h^H (\nabla f_h(y^k_h) + \nabla(\gamma h g_h)(y^k_h)) - \nabla f_H(x^k_{H,0}) + \nabla(\gamma h g_H)(x^k_{H,0})
\]

(9)

The third term in (8) is added to enforce the first order coher-
ence between a smoothed coarse objective function

\[
F_{h,\gamma h}(x^k_h) = f_h(x^k_h) + \gamma h g_h(x^k_h) + \langle v_{H,k}, x^k_h \rangle
\]

(10)

and a smoothed fine objective function \( F_{h,\gamma h} \) [12] near \( x^k_{h,0} \):

\[
\nabla F_{h,\gamma h}(x^k_{h,0}) = I_h^H \nabla F_{h,\gamma h}(y^k_h).
\]

(11)

The choice of the smoothing parameters \( \gamma_h \) and \( \gamma_H \) will
be discussed in Section 4. This condition ensures that if \( x^k_{h,m} - x^k_{h,0} \)
is a descent direction for \( F_{H,\gamma h} \) at \( x^k_{h,0} \), then \( I_h^H(x^k_{h,m} - x^k_{h,0}) \) is a descent direction for \( F_{H,\gamma_H} \) as well:

\[
(I_h^H(x^k_{h,m} - x^k_{h,0}), \nabla F_{h,\gamma h}(y^k_h)) \leq 0.
\]

(12)

According to properties of the Moreau envelope and the princi-
ple developed in [17], if \( x^k_{h,m} - x^k_{h,0} \) is a descent direction for
\( F_{H,\gamma_H} \), we obtain

\[
F_h(y^k_h + \tau_h \gamma_h \nabla f_h(y^k_h - \tau_h \nabla f_h(y^k_h)) \leq F_h(y^k_h) + \beta \gamma_h
\]

(13)
where \( \tau_{h,k} \) controls that the update is not too big and where \( \beta > 0 \) depends on \( g_h \). This ensures that \( F_h \) is decreasing up to a constant \( \beta \gamma_h \) (which can be made arbitrarily small) after a use of the coarse models. Now we show how to enforce the decrease of \( F_{H,\gamma_H} \).

**Minimization operator** \( \Phi_H \) – At the coarse level we can decide to consider either the non-smooth approximation (8) of the objective function or the smoothed version (10). Both cases lead to a decrease in \( F_{H,\gamma_H} \); indeed, taking the Moreau envelope of \( g_h \), in \( F_H(x_{k,m}) \leq F_H(x_{k,0}) \) yields \( F_{H,\gamma_H}(x_{k,m}^H) \leq F_{H,\gamma_H}(x_{k,0}^H) \). The two cases are linked by the same choice of the correction term to ensure the coherence between the two levels (9). We consider here three different strategies:

1. Gradient steps on the smoothed \( F_{H,\gamma_H} \):
   \[ \Phi_H^S = (\text{Id} - \tau_H(\nabla f_H + v_H)) \]

2. Proximal gradient steps on the non-smooth \( F_H \):
   \[ \Phi_H^F = \text{prox}_{\tau_H g_H}(\text{Id} - \tau_H(\nabla f_H + v_H)) \]

3. FISTA steps on the non-smooth \( F_H \) with the previous proximal gradient step and where \( \alpha_{H,k}^{H,L} \) follows (AD) conditions. Noted \( \Phi_{FISTA,H} \) in the following.

**Practical considerations** – Our algorithm is based on a simple construction of \( F_H \) and \( \varphi_{H,k} \), as long as the computation of the associated proximal operator has an explicit form, which is a rather reasonable assumption. Our method is sketched in Algorithm 1. The step length at both levels can be selected either by fixing a value below the threshold guaranteeing convergence, defined by the Lipschitz constants associated to the considered functions when they are known, or by a linear search to guarantee it. To ensure the convergence of the iterates, we impose at most \( p \) uses of the coarse models \( F_H \) (one use corresponds to a full V-scheme cycle), which is also recommended to significantly improve the computation time (cf. Section 4).

### 3. CONVERGENCE OF THE ITERATES

Provided that we use the coarse models a finite number of times, we can prove the convergence of the iterates to a minimizer of \( F = F_h \) and that the rate of convergence remains \( O(1/k^2) \). First, we consider the sequence of corrections from the coarse models.

**Lemma 1.** Let \( L_{f,H} \) and \( L_{f,H} \) the Lipschitz constants of \( f_h \) and \( f_{H} \), respectively. Let \( \tau_h, \tau_H \in (0, +\infty) \) the step sizes taken at fine and coarse levels, respectively. Assume that \( \tau_H < (L_{f,H})^{-1} \) and that \( \tau_h < L_{f,H}^{-1} \) and denote \( \overline{\tau}_h = \sup_k \overline{\tau}_{h,k} \). The sequence \( (c_k^h)_{k \in \mathbb{N}} \) is generated by Algorithm 1 defined by:

\[
c_k^h = \nabla f_h(y_k^H) - \nabla f_h(y_k^H) + (\tau_h)^{-1} \tau_H I_{H,0}^H(x_{k,m}^H - x_{k,0}^H)
\]

if a coarse correction is used at iteration \( k \) and \( c_k^H = 0 \) otherwise, is such that \( \sum_{k \in \mathbb{N}} \|c_k^H\| < +\infty \).

The proof of this lemma is based on the fact that if the number of coarse corrections is finite, we only need to construct bounded sequences at coarse level so that \( I_{H}^H(x_{k,m}^H - x_{k,0}^H) \) is also bounded. From this result we deduce the following theorem:

**Theorem 1.** Consider Algorithm 1, suppose that for all \( k \in \mathbb{N} \), \( \lambda_k^H \) in Eqs. (3) and (4) satisfy (AD) conditions [15]. Suppose moreover that the assumptions of Lemma 1 hold. Then:

- The sequence \( (\lambda_k^H (F_h(x_k^H) - F_h(x^*))_{k \in \mathbb{N}} \) belongs to \( \ell_{\infty}(\mathbb{N}) \).
- The sequence \( (x_k^H)_{k \in \mathbb{N}} \) given by Algorithm 1 weakly converges to a minimizer of \( F_h \).

**Proof.** We combine [15, Theorem 3.5, 4.1, and Corollary 3.8] with Lemma 1 to prove the desired result.

### 4. RESULTS

We numerically illustrate the performance of our algorithm in the context of image restoration.

**Dataset and degradation** – We consider large images of size \( 2048 \times 2048 \), yielding \( N = (2^{12})^2 \approx 4 \times 10^6 \) with \( J = 11 \). The linear degradation operator \( A_h \) is constructed with HNO [18] as a Kroenecker product with Neumann boundary conditions and we add a Gaussian noise (see the legend of Fig. 1 for details). In all tests, the regularization parameter \( \lambda_h \) was chosen by a grid search, in order to maximize the Signal-to-Noise-Ratio (SNR) of \( \hat{x} \) obtained with FISTA at convergence. Also, we initialise \( z_0 \) with the Wiener filtering of \( z \).

**Multilevel architecture** – We use a 5-levels hierarchy; from \( 2048 \times 2048 \) (\( J = 11 \)) to \( 128 \times 128 \) (indexed by \( J = 4 \)). We choose \( I_{H}^H \) as the low scale projection on a symlet wavelet.
with 10 vanishing moments and $I_h^b = \frac{1}{2}(I_h^b)^T$. We then construct $f_H$ with the blurring matrix $A_H = (I_h^b A_h I_h^b)$ (which is never used explicitly due to the properties of the Krolcnecker product [12, 18]). Thus $f_h = \frac{1}{2}||A_h x_h - z_h||^2$ and $f_H = \frac{1}{2}||A_H x - H_H z_h||^2$. The penalty term $g_h = \|W_h x_h\|_1$ is defined using a full wavelet decomposition over $J$ levels, we construct $g_H = \|W x_H\|_1$ with a decomposition over $J - 1$ to $J - 4$ levels, with $\lambda_H = \gamma_h/4$. The Moreau envelope parameter associated with $g_h$ is set to $\gamma_H = 1.1$ while $\gamma_h$ is set to 1, but both values do not seem to be critical here.

**Visual result** – We display the restored image $\hat{x}$ and the convergence curves as a function of the iterations and the CPU time for one case in Fig.1. For clarity, we only display the behaviour of the method with $\Phi_{H,FISTA}$.

**Performance assessment** – We measure $\text{Time(MMFISTA)}$, the CPU time needed to reach a threshold of 5, 2, 1, 0.1 and 0.01% of the distance $\|F_h(x_h^0) - F_h(\tilde{x})\|$, with $\tilde{x}$ computed beforehand by FISTA, and we compare it to $\text{Time(FISTA)}$. We tested the performance for several values of $m$, and among our numerous numerical experiments, $m = 5$ at the different coarse levels appears to be a good compromise whatever the noise and blur levels. We report in Tab.1, the quantity:

\[
\frac{\text{Time(MMFISTA)} - \text{Time(FISTA)}}{\text{Time(FISTA)}} \times 100, \quad (14)
\]

for $m = 5$ at every coarse levels. In this table we evaluate:

– Impact of $p$. In our numerical experiments we only consider $p = 1$ (●) or $p = 2$ (●) uses of the coarse models, performed at the beginning of the iterative process. They allow to quickly determine the low frequencies components of the solution at the fine level. The choice of $p$ depends on the desired accuracy. For rough approximation (e.g. 5%), fixing $p = 1$ is a good choice, while $p = 2$ seems better for finer approximations (e.g. thresholds set to 2%). When very precise approximation is wanted (e.g. threshold below 1%), the use of a multilevel strategy does not pay off, even if it does not degrade performance.

– Impact of noise and blur level. For all methods acceleration increases significantly as the blur gets worse. Moreover, as the noise decreases, the improvement obtained with $\Phi_{H,FISTA}$ increases, compared to others $\Phi_H$.

The main conclusion is that with a few coarse corrections, our method can significantly accelerate the achievement of coarse solution accuracies, while it stays competitive with FISTA for high precision approximations.

<table>
<thead>
<tr>
<th>Noise \ Blur</th>
<th>(a) size(blur) = $[40, 40]$, $\sigma$(blur) = 7.3</th>
<th>(b) size(blur) = $[88, 88]$, $\sigma$(blur) = 16</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\sigma = 0.01$</td>
<td></td>
<td></td>
</tr>
<tr>
<td>FISTA CPU time</td>
<td>16</td>
<td>28</td>
</tr>
<tr>
<td>$\Phi_{H,FISTA}$</td>
<td>$-20 \bullet$</td>
<td>$-22 \bullet$</td>
</tr>
<tr>
<td>$\Phi_{H,FISTA}$</td>
<td>$-19 \bullet$</td>
<td>$-19 \bullet$</td>
</tr>
<tr>
<td>$\Phi_{I,FISTA}$</td>
<td>$-31 \bullet$</td>
<td>$-32 \bullet$</td>
</tr>
</tbody>
</table>

Table 1: For each degradation : the first line of each subtable represents the computation time (in sec) needed by FISTA to reach 5, 2, 1, 0.1 and 0.01% of the distance $\|F_h(x_h^0) - F_h(\tilde{x})\|$. Then for each type of minimization algorithm at coarse level, we display the CPU time relative to FISTA (14) (in %) for the best configuration with a colored bullet : $p = 1$ (●) and $p = 2$ (●). In all cases : $m = 5$. SNR of $z : (1a)$ $11.05 \pm 0.94$ (2a) $11.03 \pm 0.93$. SNR of $x_h, 300$ computed by MMFISTA : (1a) $12.71 \pm 0.11$ (2a) $12 (2b) 10.6$. A close-up of leaves in Glacier National Park, Montana taken by Ansel Adams in the 1930s.
6. REFERENCES


