Rates of Convergence of Perturbed FISTA-based algorithms

Jean-François Aujol, Charles Dossal, Gersende Fort, Éric Moulines

Rates of Convergence of Perturbed FISTA-based algorithms

Jean-Francois Aujol*[1], Charles Dossal†[2], Gersende Fort‡[3] and Eric Moulines§[4]

[1]Institut de Mathématiques de Bordeaux, Université de Bordeaux, France
[2]Institut de Mathématiques de Toulouse, INSA and Université de Toulouse, France
[3]Institut de Mathématiques de Toulouse, CNRS and Université de Toulouse, France
[4]Centre de Mathématiques Appliquées, Ecole Polytechnique and Institut Polytechnique de Paris, France

July 14, 2019

1 Introduction

To minimize a structured convex function $F = f + g$ with f a smooth function whose gradient is $L-$Lipschitz and g a simple function whose proximal operator can be computed, a classical algorithm is the Forward-Backward (FB) algorithm also called Proximal-Gradient algorithm. The FB algorithm alternates an explicit gradient step on f and a proximal descent on g. The sequence $\{\theta_n,\ n \in \mathbb{N}\}$ built by the FB algorithm converges to a minimizer $\theta_\star$ of F and it satisfies $F(\theta_n) - \min F = O(n^{-1})$. Based on the ideas of Nesterov, FISTA proposed by Beck and Teboulle (2009) is an acceleration of FB using an extrapolation step. With this extrapolation scheme, the sequence $\{\theta_n,\ n \in \mathbb{N}\}$ satisfies $F(\theta_n) - \min F = O(n^{-2})$. In many numerical experiments FISTA ensures a better decay of the value of the functional F than FB. Nevertheless, FISTA seems to be less robust to perturbations. If the gradient of f used at each step of FB is inexact, the sequence $\{\theta_n,\ n \in \mathbb{N}\}$ converges under

*jaujol@math.u-bordeaux.fr
†Charles.Dossal@insa-toulouse.fr
‡gersende.fort@math.univ-toulouse.fr
§eric.moulines@polytechnique.edu

conditions on the perturbations, and the decay of $F(\theta_n) - \min F$ may be optimal if the error ϑ_n on the gradient and the stepsize sequence $\{\gamma_n,\ n \in \mathbb{N}\}$ satisfy conditions essentially of the form $\sum_n \gamma_n \eta_n < +\infty$ (with probability one in the case of random perturbations). In Atchade et al. (2014); Aujol and Dossal (2015); Schmidt et al. (2011); Fort et al. (2018b), the authors proved that under more restrictive assumptions on the perturbations of the gradient, the decay of $F(\theta_n) - \min F$ remains optimal and the sequence $\{\theta_n,\ n \in \mathbb{N}\}$ converges.

In this paper, the convergence of a class of inertial Forward-Backward is studied when the perturbations are both deterministic and non deterministic. Bounds on the mean and on the variance of error on the gradient are given ensuring the optimal decay of $\{F(\theta_n), n \in \mathbb{N}\}$ and the convergence of the sequence $\{\theta_n,\ n \in \mathbb{N}\}$. The stochastic perturbations setting corresponds to the case ∇f is an expectation and is estimated by Monte Carlo sampling at each step; the role of the variance of these Monte Carlo approximations on the convergence rate is also discussed in this paper.

The main contribution of this paper is to combine the stability results of Aujol and Dossal (2015) to the perturbed analysis provided in Atchade et al. (2017) (see also Atchade et al. (2014)) with an emphasis on the stochastically perturbed algorithms. The paper also weakens the conditions in Aujol and Dossal (2015) on the perturbations of the gradient, an improvement which is especially crucial in the case of random perturbations.

The paper is organized as follows. In Section 2, we define the approximate inertial Forward-Backward algorithm, FB and FISTA being two special cases. In Section 3, we recall the known results on these algorithms when the perturbations are deterministic. In Section 4, we state extensions of (Atchade et al., 2014, section 5) (see also Fort et al. (2018b)) to more general relaxations and state new results when perturbations are random. Section 5 discusses the rate of convergence for different Monte Carlo strategies. Appendix A part is dedicated to technical proofs.

2 Assumptions and Algorithm

In this section, we introduce the optimisation problem studied in this work, as well as the assumptions that we use to establish convergence results.

This paper deals with first-order methods for solving the problems:

$$\textbf{(P)}\ \mathrm{Argmin}_{\theta \in \mathbb{R}^p} F(\theta) \quad \text{or} \quad \min_{\theta \in \mathbb{R}^p} F(\theta) \qquad \text{with } F = f + g\ ,$$

when the functions f, g satisfy

H1. *The function $g : \mathbb{R}^p \to [0, +\infty]$ is convex, not identically $+\infty$, and lower semi-continuous. The function $f : \mathbb{R}^p \to \mathbb{R}$ is convex, continuously differentiable on $\mathbb{R}^p$*

and there exists a finite non-negative constant L such that, for all $\theta, \theta' \in \mathbb{R}^p$,

$$\|\nabla f(\theta) - \nabla f(\theta')\| \leq L\|\theta - \theta'\|\,,$$

where ∇f denotes the gradient of f.

We denote by Θ the domain of g: $\Theta \stackrel{\text{def}}{=} \{\theta \in \mathbb{R}^p : g(\theta) < \infty\}$.

H2. *The set $\mathcal{L} \stackrel{\text{def}}{=} \text{argmin}_{\theta\in\Theta} F(\theta)$ is a non empty subset of Θ.*

Define for any $\gamma > 0$, the proximal operator: for any $\theta \in \mathbb{R}^p$,

$$\text{Prox}_{\gamma,g}(\theta) \stackrel{\text{def}}{=} \text{Argmin}_{\tau\in\Theta}\, g(\tau) + \frac{1}{2\gamma}\|\tau - \theta\|^2, \tag{1}$$

and set for $\theta \in \mathbb{R}^p$,

$$T_{\gamma,g}(\theta) \stackrel{\text{def}}{=} \text{Prox}_{\gamma,g}\left(\theta - \gamma\nabla f(\theta)\right). \tag{2}$$

Then, the FISTA-based algorithm is given by

Input: An initial value $\theta_0 \in \Theta$, and two positive sequences $\{\gamma_n,\ n \in \mathbb{N}\}$ and $\{t_n,\ n \in \mathbb{N}\}$ satisfying

$$\gamma_n \in (0, 1/L]\,, \quad t_0 = 1, t_n \geq 1, \quad \gamma_{n+1}t_n(t_n - 1) \leq \gamma_n t_{n-1}^2. \tag{3}$$

Initialisation Set $\vartheta_0 = \theta_0$
For $n = 0, \cdots$, construct an approximation G_{n+1} of $\nabla f(\vartheta_n)$ set

$$\theta_{n+1} = \text{Prox}_{\gamma_{n+1},g}\left(\vartheta_n - \gamma_{n+1}G_{n+1}\right), \tag{4}$$

$$\alpha_{n+1} = \frac{t_n - 1}{t_{n+1}}, \tag{5}$$

$$\vartheta_{n+1} = \theta_{n+1} + \alpha_{n+1}\left(\theta_{n+1} - \theta_n\right). \tag{6}$$

Return the path $\{\theta_n, n \geq 0\}$

Some of the results below will be obtained under the following restrictive assumptions:

H3. *Θ is bounded.*

H4. *For any* $n \geq 1$,

$$\gamma_n = \gamma, \qquad t_n = (n+a-1)^d / a^d,$$

where $\gamma \in (0, 1/L]$, $d \in (0,1]$ *and*

$$\begin{cases} a > 1 & \text{if } d \in (0, 1/2)\,, \\ a > (2d)^{1/d} & \text{otherwise} \end{cases}$$

It is proved in Lemma 18 that the sequences $\{\gamma_n,\ n \in \mathbb{N}\}$ and $\{t_n,\ n \in \mathbb{N}\}$ given by H4 satisfy the condition (3).

3 Perturbed FISTA-based algorithms: rates of convergence

In this section, we improve on the known results about FISTA in the deterministic case: the results presented here adapted from Aujol and Dossal (2015) and weaken the assumptions on the perturbations. When applied to stochastic perturbations (see Section 4), this improvement is fundamental.

Define the perturbation of update scheme at each iteration, of the FISTA-based algorithm

$$\eta_{n+1} \stackrel{\text{def}}{=} G_{n+1} - \nabla f(\vartheta_n). \tag{7}$$

The following theorem is proved in Section A.2.

Theorem 1. *Assume H1 and H2. Let* $\{\theta_n,\ n \in \mathbb{N}\}$ *and* $\{\vartheta_n,\ n \in \mathbb{N}\}$ *be given by* (4) *and* (6)*, applied with positive sequences* $\{t_n,\ n \in \mathbb{N}\}$ *and* $\{\gamma_n,\ n \in \mathbb{N}\}$ *satisfying* (3)*. Set* $\overline{\Delta}_n \stackrel{\text{def}}{=} t_n \left(T_{\gamma_{n+1},g}(\vartheta_n) - \theta_n\right) + \theta_n$.

(i) If there exists $\theta_\star \in \mathcal{L}$ *such that*

$$\sum_n \gamma_{n+1}^2 t_n^2 \|\eta_{n+1}\|^2 < \infty, \qquad \sum_n \gamma_{n+1} t_n \left\langle \overline{\Delta}_n - \theta_\star, \eta_{n+1} \right\rangle < \infty, \tag{8}$$

then,

$$\sum_n \left(\gamma_n t_{n-1}^2 - \gamma_{n+1} t_n (t_n - 1)\right) \left(F(\theta_n) - \min_\Theta F\right) < \infty, \tag{9}$$

$$\sup_n \gamma_{n+1} t_n^2 \left(F(\theta_{n+1}) - \min_\Theta F\right) < \infty, \tag{10}$$

$$\sup_n \|\overline{\Delta}_n\| < \infty. \tag{11}$$

(ii) For any $\theta_\star \in \mathcal{L}$,

$$F\left(\theta_{n+1}\right) - \min_\Theta F \leq \frac{C_1 + C_{2,n} + C_{3,n}}{\gamma_{n+1} t_n^2}$$

where $C_1 \stackrel{\text{def}}{=} \gamma_1 \left(F(\theta_1) - \min_\Theta F\right) + \frac{1}{2}\|\theta_1 - \theta_\star\|^2$ *and*

$$C_{2,n} \stackrel{\text{def}}{=} \sum_{k=1}^n \gamma_{k+1}^2 t_k^2 \|\eta_{k+1}\|^2, \qquad C_{3,n} \stackrel{\text{def}}{=} -\sum_{k=1}^n \gamma_{k+1} t_k \left\langle \overline{\Delta}_k - \theta_\star, \eta_{k+1} \right\rangle.$$

The second condition in (8) can be difficult to check in practice since it may be non trivial to control the sequence $\{\overline{\Delta}_n,\ n \in \mathbb{N}\}$. It is proved in Lemma 9 that if $\sum_n \gamma_{n+1} t_n \|\eta_{n+1}\| < \infty$, then both conditions in (8) are satisfied. The property $\sum_n \gamma_{n+1} t_n \|\eta_{n+1}\| < \infty$ also implies that $C_1 + \sup_n C_{2,n} + \sup_n |C_{3,n}| < \infty$, so that $F(\theta_n) - \min_\Theta F = O(1/(\gamma_{n+1} t_n^2))$.

To optimize the decay of $F(\theta_n) - \min_\Theta F$ Nesterov proposed to choose a parameter sequence achieving the equality in (3), which corresponds for a constant step $\gamma_n = \gamma$ to $t_{n+1} = \frac{1+\sqrt{1+4t_n^2}}{2}$ and which leads to $F(\theta_n) - \min_\Theta F = O(\frac{1}{n^2})$ when the perturbations vanish. We can observe that the same decay can be achieved with $t_n = \frac{n+1-a}{a}$ with $a \geqslant 2$. It turns out that this choice may not be optimal when the serie $\sum_n \gamma_{n+1}^2 n^2 \|\eta_{n+1}\|^2$ diverges. In this case it may be better to slow down the acceleration choosing a sequence $\{t_n,\ n \in \mathbb{N}\}$ given by H4 with $d < 1$ and to average the sequence of parameters.

More precisely we have the following Corollary (see also Aujol and Dossal (2015)):

Corollary 2 (of Theorem 1)**.** *(i) If* $\lim_n \gamma_{n+1} t_n^2 = +\infty$*, then the cluster points of the sequence* $\{\theta_n,\ n \in \mathbb{N}\}$ *are in* $\mathcal{L}$*.*

(ii) If the sequences $\{t_n,\ n \in \mathbb{N}\}$ *and* $\{\gamma_n,\ n \in \mathbb{N}\}$ *are given by H4, then*

$$\sum_n n^d \left(F(\theta_n) - \min_\Theta F\right) < \infty, \qquad \sup_n n^{2d} \left(F(\theta_n) - \min_\Theta F\right) < \infty.$$

(iii) Let $\{s_n,\ n \in \mathbb{N}\}$ *and* $\{z_n,\ n \in \mathbb{N}\}$ *be defined by* $s_n \stackrel{\text{def}}{=} \sum_{k=\frac{n}{2}}^n t_k$ *and* $z_n \stackrel{\text{def}}{=} s_n^{-1} \sum_{k=\frac{n}{2}}^n t_k \theta_k$*. Then,*

$$F(z_k) - \min_\Theta F = o\left(n^{-(d+1)}\right).$$

Proof. The proof of the first two points follows from (9), (10) and Lemma 18. From the second item of the corollary, for any $\varepsilon > 0$, there exists n_0 such that for any

$n \geqslant n_0$,

$$\sum_{k=\lfloor \frac{n}{2} \rfloor}^{n} t_k \left(F(\theta_k) - \min_{\Theta} F \right) \leqslant \varepsilon.$$

Since F is convex by H1, it follows that $s_n(F(z_n) - \min F) \leqslant \varepsilon$. Then we conclude by observing that $s_n \sim C\, n^{1+d}$ for some $C > 0$. □

It turns out that such bounds can not be reached with Theroem 1 using FISTA or classical FB.

We now discuss the convergence of the iterates. The proof of the weak convergence of iterates is classical for the (exact) FB algorithm and relies on fixed point theorems; it turns out that the convergence of the sequence $\{\theta_n,\ n \in \mathbb{N}\}$ for FISTA and more generally for Nesterov acceleration scheme has been proved years after, in Chambolle and Dossal (2014) without any perturbations and in Aujol and Dossal (2015) if one considers perturbations both on the gradient and on the proximal step.

In the case the proximal operator can be computed exactly but the gradient is approximated, the following result improves on Aujol and Dossal (2015); an improvement which is especially relevant for stochastic perturbations; the proof is in Section A.3.

Theorem 3. *Assume H1 and H2. Let $\{\theta_n,\ n \in \mathbb{N}\}$ be given by (4) applied with positive sequences $\{t_n,\ n \in \mathbb{N}\}$ and $\{\gamma_n,\ n \in \mathbb{N}\}$ satisfying (3). Assume in addition that*

$$\lim_n \sum_{m=1}^{n} \sum_{k=2}^{m+1} \gamma_k \left(\prod_{i=k}^{m} \alpha_i \right) \langle \eta_k, \theta_k - \theta_\star \rangle \ \text{exists}, \tag{12}$$

$$\sum_{k \geq 1} \left(\sum_{n \geq k} \prod_{i=k}^{n} \alpha_i \right) \frac{\alpha_k + 1}{2} \left\| \theta_k - \theta_{k-1} \right\|^2 < \infty \,. \tag{13}$$

Then, for any $\theta_\star \in \mathcal{L}$, $\lim_n \|\theta_n - \theta_\star\|$ exists.

This theorem yields the following corollary. This result relies on the Opial Lemma and a complete proof can be found in (Aujol and Dossal, 2015, Theorem 4.1).

Corollary 4. *If the sequences $\{t_n,\ n \in \mathbb{N}\}$ and $\{\gamma_n,\ n \in \mathbb{N}\}$ are given by H4 and if*

$$\sum_n n^d \left\| \eta_n \right\|_2 < +\infty$$

then the sequence $\{\theta_n,\ n \in \mathbb{N}\}$ converges to a minimizer of F.

4 Case of stochastic perturbations

This section applies the previous results to the case the perturbation is stochastic. It extends earlier works (see e.g. Atchade et al. (2014); Fort et al. (2018b)) to the case $d \in (0,1)$ in H4. It is shown that d in H4 can be chosen as the decaying rate of the bias and the variance of the stochastic approximation G_n. Define the filtration

$$\mathcal{F}_n \stackrel{\text{def}}{=} \sigma\left(\theta_0, G_1, \cdots, G_n\right).$$

H 5. *The error is of the form*

$$\eta_{n+1} = \epsilon_{n+1} + \xi_{n+1}$$

where $\{\epsilon_n,\ n \in \mathbb{N}\}$ *is a martingale-increment sequence with respect to the filtration* $\{\mathcal{F}_n,\ n \in \mathbb{N}\}$*, the random sequence* $\{\xi_n,\ n \in \mathbb{N}\}$ *is* $\mathcal{F}_n$*-adapted and there exist constants* $\mathsf{a} \in [0, +\infty)$*,* $\mathsf{b} \in [0, +\infty)$ *and* $C_\epsilon, C_\xi \geq 0$ *such that*

$$\forall n \geq 1, \qquad \mathbb{E}\left[|\epsilon_{n+1}|^2 | \mathcal{F}_n\right] \leq \frac{C_\epsilon}{n^{2\mathsf{a}}} \ a.s. \qquad \mathbb{E}\left[|\xi_n|^2\right] \leq \frac{C_\xi}{n^{2\mathsf{b}}}.$$

Theorem 5. *Assume H1, H2 and H5. Let* $\{\theta_n,\ n \in \mathbb{N}\}$ *be given by* (4) *applied with the sequences* $\{t_n,\ n \in \mathbb{N}\}$ *and* $\{\gamma_n,\ n \in \mathbb{N}\}$ *given by H4. Assume in addition that*

$$C_\epsilon \sum_n n^{2(d-\mathsf{a})} + C_\xi \sum_n n^{d-\mathsf{b}} < \infty,$$

then a.s.

$$\sum_n n^d \left(F(\theta_n) - \min_\Theta F\right) < \infty, \qquad \sup_n n^{2d} \left(F(\theta_n) - \min_\Theta F\right) < \infty, \qquad \sup_n \|\overline{\Delta}_n\| < \infty.$$

Moreover, define $\{s_n,\ n \in \mathbb{N}\}$ *and* $\{z_n,\ n \in \mathbb{N}\}$ *by* $s_n \stackrel{\text{def}}{=} \sum_{k=\lfloor \frac{n}{2} \rfloor}^n k^d$ *and* $z_n \stackrel{\text{def}}{=} s_n^{-1} \sum_{k=\lfloor \frac{n}{2} \rfloor}^n k^d \theta_k$*; then,*

$$F(z_k) - \min_\Theta F = o\left(n^{-(d+1)}\right).$$

If in addition H3 holds, then

(i) $\sup_n n^d \|\theta_{n+1} - \theta_n\| < \infty$ *a.s. and* $\sum_n n^d \|\theta_n - \theta_{n-1}\|^2 < \infty$ *a.s.*

(ii) there exists a $\mathcal{L}$*-valued random variable* $\theta_\star$ *such that* $\lim_n \theta_n = \theta_\star$ *a.s.*

5 Application to Monte Carlo approximations

In this section, we apply the results of the previous section to analyze the specific case when $\nabla f(\theta)$ is an untractable expectation w.r.t. a distribution $\mathrm{d}\pi_\theta$:

$$\nabla f(\theta) = \int H(\theta, x)\ \mathrm{d}\pi_\theta(x).$$

This situation occurs especially when ∇f can be written as an expectation with respect to some target distribution: an expectation in high dimension or a distribution which is known up to a normalizing constant for example (see e.g.(Atchade et al., 2017, Sections 4 and 5) and Fort et al. (2018a)). The bounds derived in Theorem 5 can be used to control the error of the algorithm when at each step, the approximation G_{n+1} is built using Monte Carlo samples $\{X_{n+1,j}, j \geq 0\}$,

$$G_{n+1} = \frac{1}{m_{n+1}} \sum_{j=1}^{m_{n+1}} H(\vartheta_n, X_{n+1,j})$$

either exactly sampled from π_{ϑ_n} or approximating π_{ϑ_n}. In this setting the law of the random variable η_n depends on the way the Monte Carlo points are sampled, and the bias and variance of the error depend on the number of Monte Carlo points at the step n. Hence we can deduce from Theorem 5 a sampling strategy ensuring the best decay of $F(\theta_n) - \min_\theta F$.

Case of independent and identically distributed (i.i.d.) samples. The situation where G_{n+1} is computed by a usual MonteCarlo sampling using $m_n \sim n^{\mathsf{c}}(\ln n)^{\bar{\mathsf{c}}}$ i.i.d. points from $\mathrm{d}\pi_{\vartheta_n}$ corresponds to $\xi_n = 0$ (so $C_\xi = 0$) in Theorem 5 and $\mathsf{a} = \mathsf{c}/2$; here $\bar{\mathsf{c}}$ is assumed large enough for the convergence of series to hold and its value is not detailed in the discussions below.

To apply Theorem 5, we must choose c s.t. $\mathsf{c} \geq 2d + 1$ (up to a logarithmic term in the definition of m_n). Hence taking at the step n of the algorithm n^{2d+1} samples to build G_{n+1} one can ensure that $F(z_n) - \min_\theta F = o(n^{-(1+d)})$. The maximal rate of convergence is thus reached with $d = 1$, for the averaging sequence $\{z_n,\ n \in \mathbb{N}\}$ when the weights are $t_n = O(n^d)$. In Atchade et al. (2017), it is proved that the rate of convergence after n iterations of the stochastic FB algorithm (which corresponds to $d = 0$) is $O(1/n)$ for the same averaging sequence (note that in FB, $t_n = 1$) and a Monte Carlo batch size increasing as $m_n = O(n)$; our results in this paper are thus homogeneous with the case $d = 0$ adressed in Atchade et al. (2017). It was also proved in Atchade et al. (2014) that for the stochastic FISTA (which corresponds to $d = 1$), $F(\theta_n) - \min_\Theta F$ is $O(1/n^2)$ after n iterations, by choosing $m_n = O(n^3)$; our results in this paper establish a better result by showing that it is $o(1/n^2)$.

Nevertheless, for any $d \in (0,1]$, the Monte Carlo cost of this strategy is $N = O(n^{2d+2})$ samples after n iterations of the algorithm. It follows that for a Monte Carlo budget N, only $n_N = O(N^{1/(2d+2)})$ iterations can be performed and $F(z_{n_N}) - \min_\theta F = o\left(N^{\frac{-(1+d)}{2d+2}}\right) = o(N^{-1/2})$. Similar conclusions (with o replaced by O) were reached in Atchade et al. (2017) and Atchade et al. (2014) respectively for $d = 0$ and $d = 1$. Note that when the computational complexity is considered, the choice of $d \in [0,1]$ is not relevant for the rate of convergence.

Case of Markov chain Monte Carlo samples. If G_{n+1} is computed via a Markov chain Monte Carlo sampler, with n^{c} samples at iteration n, then the approximation G_{n+1} is a biased approximation of $\nabla f(\vartheta_n)$ so that we have $C_\xi \neq 0$. Under ergodicity conditions on the sampler (see e.g. (Atchade et al., 2017, Proposition 5)), the value of a in Theorem 5 can be set to $\mathsf{a} = \mathsf{c}/2$ as previously and the value of b can be set to $\mathsf{b} = \mathsf{c}$. Hence, Theorem 5 applies with $\mathsf{c} = 2d + 1$ (here again, up to logarithmic terms we do not discuss). The conclusion is thus the same as in the i.i.d. case above.

Case of variance reduction for Monte Carlo samplings. If i.i.d. samples from π_θ are available for any θ, then the control functional-based method proposed by Oates et al. (2017) applies. In that case, $C_\xi = 0$ since $\mathbb{E}\left[\eta_{n+1}|\mathcal{F}_n\right] = 0$ so that we have $\eta_{n+1} = \varepsilon_{n+1}$; and it is proved that if G_{n+1} is computed from $n^{\mathsf{c}}(\ln n)^{\bar{\mathsf{c}}}$ samples, then $2\mathsf{a} = 7\mathsf{c}/6$. Therefore, the conditions in Theorem 5 imply that $2d + 1 = 7\mathsf{c}/6$ - here again up to logarithmic terms - so that $\mathsf{c} = (12d + 6)/7$. Hence taking at the step n of the algorithm $n^{(12d+6)/7}(\ln n)^{\bar{\mathsf{c}}}$ samples to build G_{n+1} (for some $\bar{\mathsf{c}}$ correctly chosen), we have $F(z_n) - \min_\theta F = o(n^{-(1+d)})$. The same discussion as in the i.i.d. case holds.

After n iterations of the algorithm, the Monte Carlo cost is $N = O(n^{(12d+6)/7+1})$. It follows that given a Monte Carlo budget N, the number of iterations n_N depends on N in such a way that we have $F(z_{n_N}) - \min_\theta F = o(N^{-\frac{7(1+d)}{(12d+13)}})$. Roughly speaking, since ε is arbitrarily close to zero, the rate is of order $o(N^{-(7/12)(1-1/(12d+13))})$. Since $d \in (0,1]$, it is maximal with $d = 1$, reaching the value $o(N^{-14/25})$ which means a rate larger than $O(N^{-1/2})$.

On one hand, it is an excellent result: to our best knowledge, given a total amount N of Monte Carlo samples, the best known rate of convergence for Stochastic FISTA and for Stochastic FB-based methods (possibly combined with averaging strategies), was $O(N^{-1/2})$ (see Atchade et al. (2014, 2017)); it was achieved by using Monte Carlo procedures with standard variance i.e. $2\mathsf{a} = \mathsf{c}$ in H5 when η_{n+1} is computed with n^{c} Monte Carlo draws.

On the other hand, the above discussion does not take into account the compu-

tational cost of this Monte Carlo methods i.e. the computational cost of the control-functional based approximation G_{n+1} given $m_n \sim n^{\mathsf{c}}(\ln n)^{\bar{\mathsf{c}}}$ Monte Carlo draws; this technique requires matrix inversion of size equivalent to $m_n \times m_n$ for the computation of G_{n+1} (see Oates et al. (2017)).

A Detailed proofs

In this appendix, we state various results that are needed to prove the theorems stated in Section 3 and Section 4. Set

$$\overline{F}(\theta) \stackrel{\text{def}}{=} F(\theta) - \min_{\Theta} F, \tag{14}$$

$$\Delta_{n+1} \stackrel{\text{def}}{=} t_n(\theta_{n+1} - \theta_n) + \theta_n, \tag{15}$$

$$\overline{\Delta}_n \stackrel{\text{def}}{=} t_n \left(T_{\gamma_{n+1},g}(\vartheta_n) - \theta_n\right) + \theta_n. \tag{16}$$

Note that $\Delta_{n+1} \in \mathcal{F}_{n+1}$ and $\overline{\Delta}_n \in \mathcal{F}_n$.

A.1 Intermediate results

Lemma 6 shows that, in the stochastic case, the iterated θ_n and ϑ_n are bounded with probability one under the assumptions H3 and H4.

Lemma 6. *Assume H3 and H4. Then there exists a constant C such that*

$$\mathbb{P}(\sup_n |\theta_n| \le C) = 1, \qquad \mathbb{P}(\sup_n |\vartheta_n| \le C) = 1.$$

Proof. By definition of the prox-operator and by H3, $\theta_n \in \Theta$ and this set is bounded. Furthermore, by H4, the sequence $\{t_n,\ n \in \mathbb{N}\}$ is increasing, so that $0 \le t_{n-1} - 1 \le t_n$. This implies that $|\vartheta_n| \le |\theta_n| + |\theta_n - \theta_{n-1}|$, and the proof is concluded using again H3 and $\theta_j \in \Theta$. □

Lemma controls the difference between Δ_n and $\overline{\Delta}_n$ (see resp. (15) and (16)) as a function of the perturbation η_{n+1} and of the design parameters t_n, γ_n.

Lemma 7. *Assume H1. Let $\{\theta_n,\ n \in \mathbb{N}\}$ and $\{\vartheta_n,\ n \in \mathbb{N}\}$ be given by (4) and (6), applied with positive sequences $\{t_n,\ n \in \mathbb{N}\}$ and $\{\gamma_n,\ n \in \mathbb{N}\}$ satisfying (3). Then*

$$\|\Delta_{n+1} - \overline{\Delta}_n\| \le t_n \gamma_{n+1} \|\eta_{n+1}\|.$$

Proof. We have $\Delta_{n+1} - \overline{\Delta}_n = t_n \left(\theta_{n+1} - T_{\gamma_{n+1},g}(\vartheta_n)\right)$. Furthermore by definition of θ_{n+1} and $T_{\gamma,g}(\theta)$, we have

$$\theta_{n+1} - T_{\gamma_{n+1},g}(\vartheta_n) = \mathrm{Prox}_{\gamma_{n+1},g}\left(\vartheta_n - \gamma_{n+1} G_{n+1}\right) - \mathrm{Prox}_{\gamma_{n+1},g}\left(\vartheta_n - \gamma_{n+1} \nabla f(\vartheta_n)\right).$$

The proof is concluded upon noting that under H1, $\theta \mapsto \mathrm{Prox}_{\gamma,g}(\theta)$ is 1-Lipschitz (see e.g. (Bauschke and Combettes, 2011, Proposition 12.26)). □

The following lemma is a key building block of the studies since it establishes a Lyapunov-type inequality. Note that in the case $\eta_n = 0$ (no perturbations), there is a strict decay of the sequence of Lyapunov functions.

Lemma 8. *Assume H1 and H2. Let* $\{\theta_n,\ n \in \mathbb{N}\}$ *and* $\{\vartheta_n,\ n \in \mathbb{N}\}$ *be given by* (4) *and* (6)*, applied with positive sequences* $\{t_n,\ n \in \mathbb{N}\}$ *and* $\{\gamma_n,\ n \in \mathbb{N}\}$ *satisfying* (3)*. For any minimizer* $\theta_\star \in \mathcal{L}$*, any* $j \geq 1$*,*

$$\begin{aligned}
&\left(\gamma_j t_{j-1}^2 - \gamma_{j+1} t_j (t_j - 1)\right) \overline{F}(\theta_j) + \gamma_{j+1} t_j^2 \overline{F}(\theta_{j+1}) + \frac{1}{2}\|\Delta_{j+1} - \theta_\star\|^2 \\
&\quad \leq \gamma_j t_{j-1}^2 \overline{F}(\theta_j) + \frac{1}{2}\|\Delta_j - \theta_\star\|^2 - \gamma_{j+1} t_j \left\langle \Delta_{j+1} - \theta_\star, \eta_{j+1} \right\rangle \qquad (17) \\
&\quad \leq \gamma_j t_{j-1}^2 \overline{F}(\theta_j) + \frac{1}{2}\|\Delta_j - \theta_\star\|^2 + \gamma_{j+1}^2 t_j^2 \|\eta_{j+1}\|^2 - \gamma_{j+1} t_j \left\langle \bar{\Delta}_j - \theta_\star, \eta_{j+1} \right\rangle . \qquad (18)
\end{aligned}$$

where $\{\eta_n,\ n \in \mathbb{N}\}$*,* $\{\Delta_n,\ n \in \mathbb{N}\}$ *and* $\{\overline{\Delta}_n,\ n \in \mathbb{N}\}$ *are given by* (7)*,* (15) *and* (16)*.*

Proof. Let $j \geq 1$. We first apply Lemma 16 with $\vartheta \leftarrow \theta_j$, $\xi \leftarrow \vartheta_j$, $\theta \leftarrow \vartheta_j - \gamma_{j+1} G_{j+1}$ and $\gamma \leftarrow \gamma_{j+1}$ to get

$$2\gamma_{j+1}\overline{F}(\theta_{j+1}) \leq 2\gamma_{j+1}\overline{F}(\theta_j) + \|\theta_j - \vartheta_j\|^2 - \|\theta_{j+1} - \theta_j\|^2 - 2\gamma_{j+1} \left\langle \theta_{j+1} - \theta_j, \eta_{j+1} \right\rangle .$$

We apply again Lemma 16 with $\vartheta \leftarrow \theta_\star$ to get

$$2\gamma_{j+1}\overline{F}(\theta_{j+1}) \leq \|\theta_\star - \vartheta_j\|^2 - \|\theta_{j+1} - \theta_\star\|^2 - 2\gamma_{j+1} \left\langle \theta_{j+1} - \theta_\star, \eta_{j+1} \right\rangle .$$

We now compute a combination of these two inequalities with coefficients $t_j(t_j - 1)$ and t_j. This yields

$$\begin{aligned}
&2\gamma_{j+1} t_j^2 \overline{F}(\theta_{j+1}) + t_j(t_j - 1)\|\theta_{j+1} - \theta_j\|^2 + t_j\|\theta_{j+1} - \theta_\star\|^2 \\
&\quad \leq 2t_j(t_j - 1)\gamma_{j+1}\overline{F}(\theta_j) + t_j(t_j - 1)\|\theta_j - \vartheta_j\|^2 + t_j\|\vartheta_j - \theta_\star\|^2 \\
&\quad - 2\gamma_{j+1} t_j \left\langle \Delta_{j+1} - \theta_\star, \eta_{j+1} \right\rangle .
\end{aligned}$$

Then, by using the definition of ϑ_j and Δ_{j+1}, we have

$$\begin{aligned}
t_j(t_j - 1)\|\theta_{j+1} - \theta_j\|^2 + t_j\|\theta_{j+1} - \theta_\star\|^2 &= \|\Delta_{j+1} - \theta_\star\|^2 + (t_j - 1)\|\theta_j - \theta_\star\|^2, \\
t_j(t_j - 1)\|\theta_j - \vartheta_j\|^2 + t_j\|\vartheta_j - \theta_\star\|^2 &= \|\Delta_j - \theta_\star\|^2 + (t_j - 1)\|\theta_j - \theta_\star\|^2.
\end{aligned}$$

This yields

$$\begin{aligned}
2\gamma_{j+1} t_j^2 \overline{F}(\theta_{j+1}) + \|\Delta_{j+1} - \theta_\star\|^2 \leq 2\gamma_j t_{j-1}^2 \overline{F}(\theta_j) + \|\Delta_j - \theta_\star\|^2 - 2\gamma_{j+1} t_j \left\langle \Delta_{j+1} - \theta_\star, \eta_{j+1} \right\rangle \\
- 2\left(\gamma_j t_{j-1}^2 - \gamma_{j+1} t_j (t_j - 1)\right) \overline{F}(\theta_j) .
\end{aligned}$$

This concludes the proof. □

By using the Lyapunov-type inequalities, we are able to show that the quantities Δ_n and $\overline{\Delta}_n$ are uniformly bounded in n, under conditions on the cumulated errors.

Lemma 9. *Assume H1 and H2. Let* $\{\theta_n,\ n\in\mathbb{N}\}$ *and* $\{\vartheta_n,\ n\in\mathbb{N}\}$ *be given by* (4) *and* (6)*, applied with positive sequences* $\{t_n,\ n\in\mathbb{N}\}$ *and* $\{\gamma_n,\ n\in\mathbb{N}\}$ *satisfying* (3)*. If* $\sum_n \gamma_{n+1}t_n\|\eta_{n+1}\|<\infty$*, then*

$$\sup_n \|\Delta_n\| + \sup_n \|\overline{\Delta}_n\| < \infty.$$

Proof. By iterating (17) and since $\overline{F}\geq 0$, we have for any $\theta_\star\in\mathcal{L}$,

$$\frac{1}{2}\|\Delta_{j+1}-\theta_\star\|^2 \leq \frac{1}{2}\|\Delta_1-\theta_\star\|^2 + \gamma_1 t_0^2\overline{F}(\theta_1) - \sum_{k=1}^{j}\gamma_{k+1}t_k\left\langle \Delta_{k+1}-\theta_\star,\eta_{k+1}\right\rangle.$$

We then conclude by Lemma 19 and Lemma 7. □

Lemma 10. *Assume H1 and H2. Let* $\{\theta_n,\ n\in\mathbb{N}\}$ *and* $\{\vartheta_n,\ n\in\mathbb{N}\}$ *be given by* (4) *and* (6)*, applied with positive sequences* $\{t_n,\ n\in\mathbb{N}\}$ *and* $\{\gamma_n,\ n\in\mathbb{N}\}$ *satisfying* (3)*. Then for any* $n\geq 2$*,*

$$\begin{aligned}&2\gamma_{n+1}t_n(t_n-1)\overline{F}(\theta_{n+1}) + t_n(t_n-1)\|\theta_{n+1}-\theta_n\|^2\\ &+\sum_{k=1}^{n}\frac{t_{k-1}-1}{t_k}\left(t_k+t_{k-1}-1\right)\|\theta_k-\theta_{k-1}\|^2 \leq 2\sum_{k=1}^{n}\gamma_k t_{k-1}\overline{F}(\theta_k) + \sum_{k=1}^{n}t_k(t_k-1)\Xi_{k+1}\end{aligned}$$

where

$$\Xi_{k+1} \stackrel{\text{def}}{=} 2\gamma_{k+1}^2\|\eta_{k+1}\|^2 - 2t_k^{-1}\gamma_{k+1}\left\langle\overline{\Delta}_k-\theta_k,\eta_{k+1}\right\rangle\ . \tag{19}$$

Proof. Set $\tilde{\Xi}_{n+1}\stackrel{\text{def}}{=} -2\gamma_{n+1}\left\langle\theta_{n+1}-\theta_n,\eta_{n+1}\right\rangle$. We apply Lemma 16 with $\theta\leftarrow\vartheta_n-\gamma_{n+1}G_{n+1}$, $\vartheta\leftarrow\theta_n$, $\xi\leftarrow\vartheta_n$ and $\gamma\leftarrow\gamma_{n+1}$. This yields for any $n\geq 1$,

$$2\gamma_{n+1}F\left(\theta_{n+1}\right) + \|\theta_{n+1}-\theta_n\|^2 \leq 2\gamma_{n+1}F(\theta_n) + \|\theta_n-\vartheta_n\|^2 + \tilde{\Xi}_{n+1}\ .$$

By definition of ϑ_n, we have $\|\vartheta_n-\theta_n\|^2=\alpha_n^2\|\theta_n-\theta_{n-1}\|^2$. Hence,

$$2\gamma_{n+1}F\left(\theta_{n+1}\right) + \|\theta_{n+1}-\theta_n\|^2 \leq 2\gamma_{n+1}F(\theta_n) + \alpha_n^2\|\theta_n-\theta_{n-1}\|^2 + \tilde{\Xi}_{n+1}\ ,$$

or equivalently,

$$\|\theta_{n+1}-\theta_n\|^2 - \alpha_n^2\|\theta_n-\theta_{n-1}\|^2 \leq 2\gamma_{n+1}\left(\overline{F}(\theta_n)-\overline{F}(\theta_{n+1})\right) + \tilde{\Xi}_{n+1}\ .$$

We multiply both sides by $t_n(t_n-1)$ and sum from $k=1$ to $k=n$; we obtain on the LHS by using $t_k\alpha_k = t_{k-1}-1$ and $\alpha_1 = 0$,

$$\sum_{k=1}^{n} t_k(t_k-1)\left(\|\theta_{k+1}-\theta_k\|^2 - \alpha_k^2\|\theta_k-\theta_{k-1}\|^2\right)$$
$$= t_n(t_n-1)\|\theta_{n+1}-\theta_n\|^2 + \sum_{k=1}^{n}\frac{t_{k-1}-1}{t_k}\left(t_k+t_{k-1}-1\right)\|\theta_k-\theta_{k-1}\|^2 .$$

On the RHS, we have

$$\begin{aligned}
&2\sum_{k=1}^{n}\gamma_{k+1}t_k(t_k-1)\{\overline{F}(\theta_k)-\overline{F}(\theta_{k+1})\} + \sum_{k=1}^{n} t_k(t_k-1)\tilde{\Xi}_{k+1}\\
&\le 2\sum_{k=1}^{n}\{\gamma_{k+1}t_k(t_k-1)-\gamma_k t_{k-1}(t_{k-1}-1)\}\overline{F}(\theta_k)\\
&- 2\gamma_{n+1}t_n(t_n-1)\overline{F}(\theta_{n+1}) + \sum_{k=2}^{n} t_k(t_k-1)\tilde{\Xi}_{k+1}\\
&\le 2\sum_{k=1}^{n}\gamma_k t_{k-1}\overline{F}(\theta_k) + \sum_{k=2}^{n} t_k(t_k-1)\tilde{\Xi}_{k+1} - 2\gamma_{n+1}t_n(t_n-1)\overline{F}(\theta_{n+1})
\end{aligned}$$

where in the last inequality, we used (3). We now compute an upper bound of $\tilde{\Xi}_{n+1}$. We have

$$\theta_{n+1}-\theta_n = \theta_{n+1} - T_{\gamma_{n+1},g}(\vartheta_n) + T_{\gamma_{n+1},g}(\vartheta_n) - \theta_n .$$

Since $\mathrm{Prox}_{\gamma,g}$ is 1-Lipschitz, $\|\theta_{n+1} - T_{\gamma_{n+1},g}(\vartheta_n)\| \le \gamma_{n+1}\|\eta_{n+1}\|$; note also that $\overline{\Delta}_n - \theta_n = t_n\left(T_{\gamma_{n+1},g}(\vartheta_n)-\theta_n\right)$. This yields $\tilde{\Xi}_{n+1} \le \Xi_{n+1}$ and concludes the proof.

□

Lemma 11. *Assume H1 and H2. Let $\{\theta_n,\ n\in\mathbb{N}\}$ and $\{\vartheta_n,\ n\in\mathbb{N}\}$ be given by (4) and (6), applied with positive sequences $\{t_n,\ n\in\mathbb{N}\}$ and $\{\gamma_n,\ n\in\mathbb{N}\}$ satisfying (3). Then for any $n\ge 1$ and any $\theta_\star\in\mathcal{L}$,*

$$\|\theta_{n+1}-\theta_\star\|^2 \le \|\theta_n-\theta_\star\|^2 - \sum_{k=2}^{n+1}\left(\prod_{j=k}^{n}\alpha_j\right)\gamma_k\left(F(\theta_k)-\min_{\Theta}F\right) + \sum_{k=2}^{n+1}\left(\prod_{j=k}^{n}\alpha_j\right)B_k ,$$

where

$$B_k \stackrel{\mathrm{def}}{=} \alpha_{k-1}\frac{\alpha_{k-1}+1}{2}\|\theta_{k-1}-\theta_{k-2}\|^2 - \gamma_k\left\langle\eta_k,\theta_k-\theta_\star\right\rangle .$$

By convention, $\prod_{k=n+1}^{n}\alpha_k = 1$.

Proof. Let $\theta_\star \in \mathcal{L}$. Apply Lemma 16 with $\xi \leftarrow \vartheta_n$, $\theta \leftarrow \vartheta_n - \gamma_{n+1} G_{n+1}$, $\vartheta \leftarrow \theta_\star$ and $\gamma \leftarrow \gamma_{n+1}$. This yields

$$\|\theta_{n+1} - \theta_\star\|^2 \leq \|\vartheta_n - \theta_\star\|^2 - 2\gamma_{n+1}\overline{F}(\theta_{n+1}) + 2\gamma_{n+1} \langle \theta_{n+1} - \theta_\star, \eta_{n+1} \rangle \ .$$

By definition of ϑ_n and by using $2\langle a, b\rangle = \|a\|^2 + \|b\|^2 - \|a-b\|^2$, we have

$$\begin{aligned}\|\vartheta_n - \theta_\star\|^2 &= \|\theta_n - \theta_\star\|^2 + \alpha_n^2 \|\theta_n - \theta_{n-1}\|^2 + 2\alpha_n \langle \theta_n - \theta_\star, \theta_n - \theta_{n-1} \rangle \\ &= \|\theta_n - \theta_\star\|^2 + \alpha_n(1+\alpha_n)\|\theta_n - \theta_{n-1}\|^2 + \alpha_n \left(\|\theta_n - \theta_\star\|^2 - \|\theta_{n-1} - \theta_\star\|^2 \right).\end{aligned}$$

This yields

$$\Phi_{n+1} - \Phi_n \leq \alpha_n \left(\Phi_n - \Phi_{n-1} \right) + \left(B_{n+1} - \gamma_{n+1}\overline{F}(\theta_{n+1}) \right) \ ,$$

where $\Phi_n \stackrel{\text{def}}{=} \|\theta_n - \theta_\star\|^2/2$. By iterating (upon noting that $\alpha_n \geq 0$), we obtain

$$\begin{aligned}\Phi_{n+1} - \Phi_n &\leq \left(\prod_{j=1}^{n} \alpha_j \right) (\Phi_1 - \Phi_0) + \sum_{k=2}^{n+1} \left(\prod_{j=k}^{n} \alpha_j \right) \left(B_k - \gamma_k \overline{F}(\theta_k) \right) \\ &\leq \sum_{k=2}^{n+1} \left(\prod_{j=k}^{n} \alpha_j \right) \left(B_k - \gamma_k \overline{F}(\theta_k) \right) \ ,\end{aligned}$$

since $\alpha_1 = 0$. This concludes the proof. □

Proposition 12. *Assume H1, H2 and H5. Let* $\{\theta_n,\ n \in \mathbb{N}\}$ *and* $\{\vartheta_n,\ n \in \mathbb{N}\}$ *be given by* (4) *and* (6) *applied with the positive sequences* $\{t_n,\ n \in \mathbb{N}\}$ *and* $\{\gamma_n,\ n \in \mathbb{N}\}$ *given by H4. Assume also that*

$$C_\epsilon \sum_n n^{2(d-\mathsf{a})} + C_\xi \sum_n n^{d-\mathsf{b}} < \infty.$$

Then

$$\sup_n \mathbb{E}\left[\|\overline{\Delta}_n\|^2 \right] < \infty, \tag{20}$$

Proof. Let $\theta_\star \in \mathcal{L}$. Iterating (18) yields for any $n \geq 1$,

$$\frac{1}{2}\|\Delta_{n+1} - \theta_\star\|^2 \leq \gamma_1 \overline{F}(\theta_1) + \frac{1}{2}\|\Delta_1 - \theta_\star\|^2 + \sum_{j=1}^{n} \gamma_{j+1}^2 t_j^2 \|\eta_{j+1}\|^2 - \sum_{j=1}^{n} \gamma_{j+1} t_j \left\langle \overline{\Delta}_j - \theta_\star, \eta_{j+1} \right\rangle \ .$$

By Lemma 7, (18) and the inequality $(a+b)^2 \leq 2a^2 + 2b^2$, we have

$$\frac{1}{4}\|\overline{\Delta}_n - \theta_\star\|^2 \leq \gamma_1 \overline{F}(\theta_1) + \frac{1}{2}\|\Delta_1 - \theta_\star\|^2 + \sum_{j=1}^{n} \gamma_{j+1}^2 t_j^2 \|\eta_{j+1}\|^2 + \frac{1}{2}\gamma_{n+1}^2 t_n^2 \|\eta_{n+1}\|^2 - \sum_{j=1}^{n} \gamma_{j+1} t_j \left\langle \overline{\Delta}_j - \theta_\star, \eta_{j+1} \right\rangle \ .$$

Computing the expectation and applying the Cauchy-Schwarz inequality yield

$$\frac{1}{4}\|\overline{\Delta}_n - \theta_\star\|_2^2 \leq \gamma_1 \mathbb{E}\left[\overline{F}(\theta_1)\right] + \frac{1}{2}\mathbb{E}\left[\|\Delta_1 - \theta_\star\|^2\right] + \sum_{j=1}^{n} \gamma_{j+1}^2 t_j^2 \mathbb{E}\left[\|\eta_{j+1}\|^2\right] + \frac{1}{2}\gamma_{n+1}^2 t_n^2 \mathbb{E}\left[\|\eta_{n+1}\|^2\right] + \sum_{j=1}^{n} \gamma_{j+1} t_j \|\overline{\Delta}_j - \theta_\star\|_2 \ \|\mathbb{E}\left[\xi_{j+1}|\mathcal{F}_j\right]\|_2 \ . \quad (21)$$

where for a vector-valued r.v. U, $\|U\|_2 \stackrel{\text{def}}{=} \sqrt{\mathbb{E}\left[\|U\|^2\right]}$. We then conclude by Lemma 19, applied with $u_n^2 \leftarrow \|\overline{\Delta}_n - \theta_\star\|_2^2$, and $e_k \leftarrow 4\gamma_{k+1} t_k \ \|\mathbb{E}\left[\xi_{k+1}|\mathcal{F}_k\right]\|_2$. □

Lemma 13. *Assume H1, H2, H3 and H5. Let $\{\theta_n,\ n \in \mathbb{N}\}$ and $\{\vartheta_n,\ n \in \mathbb{N}\}$ be given by (4) and (6), applied with the positive sequences $\{t_n,\ n \in \mathbb{N}\}$ and $\{\gamma_n,\ n \in \mathbb{N}\}$ given by H4. Assume in addition that*

$$C_\epsilon \sum_n n^{2(d-\mathsf{a})} + C_\xi \sum_n n^{d-\mathsf{b}} < \infty. \quad (22)$$

Let $\{\tau_n,\ n \in \mathbb{N}\}$ be a $\mathbb{R}^p$-valued random sequence which is $\mathcal{F}_n$-adapted and such that $\sup_n \|\tau_n\| < \infty$. Then a.s.

$$\sum_k \gamma_{k+1}^2 t_k^2 \|\eta_{k+1}\|^2 < \infty, \qquad \limsup_n \left| \sum_{k=1}^{n} \gamma_{k+1} t_k \left\langle \tau_k, \eta_{k+1} \right\rangle \right| < \infty \ .$$

Proof. By the conditional Borel-Cantelli lemma (see e.g. (Chen, 1978, Theorem 1)),

$$\sum_{k \geq 1} \gamma_{k+1}^2 t_k^2 \mathbb{E}\left[\|\eta_{k+1}\|^2 | \mathcal{F}_k\right] < \infty \text{ a.s. } \Longrightarrow \sum_{k \geq 1} \gamma_{k+1}^2 t_k^2 \|\eta_{k+1}\|^2 < \infty \text{ a.s.}$$

The sufficient condition holds true by H4, H5 and (22). We write $\langle \tau_k, \eta_{k+1} \rangle = \langle \tau_k, \xi_{k+1} \rangle + \langle \tau_k, \epsilon_{k+1} \rangle$. By H5 and (22)

$$\sum_k \gamma_{k+1} t_k \|\xi_k\| < \infty \text{ a.s. } ;$$

hence, the sum $\sum_k \gamma_{k+1} t_k \langle \tau_k, \xi_k \rangle$ exists a.s. Since $\tau_k \in \mathcal{F}_k$, the term $\langle \tau_k, \epsilon_{k+1} \rangle$ is a martingale increment. Since

$$\left(\sup_n \|\tau_n\| \right)^2 \sum_k \gamma_{k+1}^2 t_k^2 \mathbb{E}\left[\|\epsilon_{k+1}\|^2 | \mathcal{F}_k \right] < \infty \text{ a.s.}$$

by H5 and (22), then (Hall and Heyde, 1980, Theorem 17) implies that the sum $\sum_k \gamma_{k+1} t_k \langle \tau_k, \epsilon_{k+1} \rangle$ exists a.s. This concludes the proof. □

Proposition 14. *Assume H1, H2, H3 and H5. Let* $\{\theta_n,\ n \in \mathbb{N}\}$ *and* $\{\vartheta_n,\ n \in \mathbb{N}\}$ *be given by* (4) *and* (6)*, applied with the positive sequences* $\{t_n,\ n \in \mathbb{N}\}$ *and* $\{\gamma_n,\ n \in \mathbb{N}\}$ *given by H4. Assume in addition that*

$$C_\epsilon \sum_n n^{2(d-\mathsf{a})} + C_\xi \sum_n n^{d-\mathsf{b}} < \infty.$$

Then a.s.

$$\sup_n n^{2d} \overline{F}(\theta_n) < \infty, \qquad \sum_n n^d \overline{F}(\theta_n) < \infty, \tag{23}$$

$$\sum_{n \geq 1} n^d \|\theta_n - \theta_{n-1}\|^2 < \infty \quad \textit{and} \quad \sup_n n^d \|\theta_{n+1} - \theta_n\| < \infty\ . \tag{24}$$

Furthermore, the condition (13) *holds a.s.*

Proof. We apply Lemma 13 with $\tau_k \leftarrow \overline{\Delta}_k - \theta_\star$. Note that by Theorem 5, we have $\sup_n \|\overline{\Delta}_n\| < \infty$ a.s. which implies, by H3, $\sup_n \|\bar{\Delta}_n - \theta_n\| < \infty$ a.s. Lemma 13 yields a.s.

$$\sum_k \gamma_{k+1}^2 t_k^2 \|\eta_{k+1}\|^2 < \infty, \qquad \limsup_n \left| \sum_{k=1}^n \gamma_{k+1} t_k \left\langle \overline{\Delta}_k - \theta_\star, \eta_{k+1} \right\rangle \right| < \infty.$$

This result, combined with Lemma 8 and Lemma 17 applied with

$$v_{j+1} \leftarrow \gamma_{j+1} t_j^2 \overline{F}(\theta_{j+1}) + \frac{1}{2} \|\Delta_{j+1} - \theta_\star\|^2,$$
$$\chi_j \leftarrow \left(\gamma_j t_{j-1}^2 - \gamma_{j+1} t_j (t_j - 1) \right) \overline{F}(\theta_j),$$

imply that $\sum_k \chi_k$ exists a.s. and $\lim_n v_n$ exists. This yields, by using Lemma 18, $\sum_k \gamma_k t_{k-1} \overline{F}(\theta_k) < \infty$ a.s. and $\sup_n \gamma_{n+1} t_n^2 \overline{F}(\theta_n) < \infty$. We obtain (23) by Lemma 18.

We apply again Lemma 13 with $\tau_k \leftarrow \overline{\Delta}_k - \theta_k$. Lemma 13 implies that $\sup_n |\sum_{k=2}^n t_k (t_k - 1) \Xi_{k+1}|$ exists a.s. where $\{\Xi_n,\ n \in \mathbb{N}\}$ is given by Lemma 10. The proof of (23) is concluded by Lemma 10 and Lemma 18.

The last claim now follows from Lemma 18 and the bound $(\alpha_n + 1)/2 \leq 1$. □

Proposition 15. *Assume H1, H2, H3, H4 and H5. Assume in addition that*

$$C_\epsilon \sum_n n^{2(d-\mathsf{a})} + C_\xi \sum_n n^{d-\mathsf{b}} < \infty. \tag{25}$$

Then the condition (12) *holds a.s.*

Proof. Throughout the proof, set

$$\beta_{k,m} \stackrel{\text{def}}{=} \prod_{i=k}^{m} \alpha_i \ , \qquad B_{k,n} \stackrel{\text{def}}{=} \sum_{m=k}^{n} \beta_{k,m} \qquad B_{k,\infty} \stackrel{\text{def}}{=} \sum_{m\geq k} \beta_{k,m} \ .$$

By Lemma 18, we have

$$\sup_k t_k^{-1} B_{k,\infty} < \infty \ . \tag{26}$$

We write for any $k \geq 2$,

$$\begin{aligned}
\langle \eta_k, \theta_k - \theta_\star \rangle &= \langle \eta_k, \theta_k - T_{\gamma_k,g}(\vartheta_{k-1}) \rangle + \langle \eta_k, T_{\gamma_k,g}(\vartheta_{k-1}) - \theta_\star \rangle \\
&= \langle \eta_k, \theta_k - T_{\gamma_k,g}(\vartheta_{k-1}) \rangle + \langle \xi_k, T_{\gamma_k,g}(\vartheta_{k-1}) - \theta_\star \rangle \\
&\quad + \langle \epsilon_k, T_{\gamma_k,g}(\vartheta_{k-1}) - \theta_\star \rangle \ .
\end{aligned}$$

Since $\mathrm{Prox}_{\gamma,g}$ is 1-Lipschitz, we have $\|\theta_k - T_{\gamma_k,g}(\vartheta_{k-1})\| \leq \gamma_k \|\eta_k\|$ so that it holds

$$\sum_{m\geq 2} \sum_{k=2}^{m} \gamma_k \left| \langle \eta_k, \theta_k - T_{\gamma_k,g}(\vartheta_{k-1}) \rangle \right| \beta_{k,m} \leq \sum_{k\geq 2} \gamma_k^2 \|\eta_k\|^2 \, B_{k,\infty} \ .$$

By (26), H5, the assumption (25) and the conditional Borel-Cantelli lemma (see e.g. (Chen, 1978, Theorem 1)), the RHS is finite a.s.

By Fubini again, the equality $T_{\gamma,g}(\theta_\star) = \theta_\star$ and the 1-Lipschitz property of $T_{\gamma,g}$ (see e.g. (Atchade et al., 2017, Lemma 9)), it holds

$$\sum_{m\geq 2} \sum_{k=2}^{m} \gamma_k \left| \langle \xi_k, T_{\gamma_k,g}(\vartheta_{k-1}) - \theta_\star \rangle \right| \beta_{k,m} \leq \sum_{k\geq 2} \gamma_k \|\xi_k\| \|\vartheta_{k-1} - \theta_\star\| B_{k,\infty} \ ,$$

and the RHS is finite a.s. by (26), (25) and Lemma 6. We now consider the last term; set

$$\Psi_k \stackrel{\text{def}}{=} \langle \epsilon_k, T_{\gamma_k,g}(\theta_k) - \theta_\star \rangle \ .$$

We have

$$\sum_{m=2}^{n} \sum_{k=2}^{m} \gamma_k \Psi_k \beta_{k,m} = \sum_{k=2}^{n} \gamma_k \Psi_k B_{k,n} = \sum_{k=2}^{n} \gamma_k \Psi_k B_{k,\infty} \frac{B_{k,n}}{B_{k,\infty}} \ . \tag{27}$$

Upon noting that $\mathbb{E}\left[\Psi_k|\mathcal{F}_{k-1}\right] = 0$, $\{\gamma_k B_{k,\infty}\Psi_k, k \geq 0\}$ is a martingale-increment sequence. We have

$$\sum_k \gamma_k^2 B_{k,\infty}^2 \, \mathbb{E}\left[\|\Psi_k\|^2|\mathcal{F}_{k-1}\right] \leq \left(\sup_n \|\theta_n - \theta_\star\|^2\right) \sum_k \gamma_k^2 B_{k,\infty}^2 \, \mathbb{E}\left[\|\epsilon_k\|^2|\mathcal{F}_{k-1}\right] \ ,$$

and the RHS is finite a.s. under (26), H3, H5, and (25). Therefore, $\lim_n \mathcal{S}_n$ exists a.s. where $\mathcal{S}_n \stackrel{\text{def}}{=} \sum_{k=2}^n \gamma_k B_{k,\infty} \Psi_k$; by convention, $\mathcal{S}_1 = 0$. By the Abel's transform and (27), it holds

$$\begin{aligned}\sum_{m=2}^n \sum_{k=2}^m \gamma_k \Psi_k \beta_{k,m} &= \sum_{k=2}^n \left(\mathcal{S}_k - \mathcal{S}_{k-1}\right) \frac{B_{k,n}}{B_{k,\infty}} \\ &= \sum_{k=2}^{n-1} \left(\frac{B_{k,n}}{B_{k,\infty}} - \frac{B_{k+1,n}}{B_{k+1,\infty}}\right) \mathcal{S}_k + \mathcal{S}_n \frac{B_{n,n}}{B_{n,\infty}} \ .\end{aligned}$$

Since $\sup_n |\mathcal{S}_n| < \infty$ a.s. and $\sup_{n,\ell} B_{n,\ell}/B_{n,\infty} \leq 1$, it is sufficient to prove that $\lim_n \sum_{k=2}^{n-1} \left|\frac{B_{k,n}}{B_{k,\infty}} - \frac{B_{k+1,n}}{B_{k+1,\infty}}\right| < \infty$. Since $\beta_{k,m} = \alpha_k \beta_{k+1,m}$, we have

$$B_{k,n} = \alpha_k + \alpha_k B_{k+1,n} \qquad B_{k,\infty} = \alpha_k + \alpha_k B_{k+1,\infty} \ .$$

This yields, using $B_{k+1,\infty} - B_{k+1,n} = \alpha_{k+1} \cdots \alpha_n B_{n+1,\infty}$

$$\frac{B_{k,n}}{B_{k,\infty}} - \frac{B_{k+1,n}}{B_{k+1,\infty}} = \alpha_k \alpha_{k+1} \cdots \alpha_n \frac{B_{n+1,\infty}}{B_{k,\infty} B_{k+1,\infty}} \geq 0 \ .$$

Hence,

$$\sum_{k=2}^{n-1} \left|\frac{B_{k,n}}{B_{k,\infty}} - \frac{B_{k+1,n}}{B_{k+1,\infty}}\right| = \sum_{k=2}^{n-1} \left(\frac{B_{k,n}}{B_{k,\infty}} - \frac{B_{k+1,n}}{B_{k+1,\infty}}\right) \leq \frac{B_{2,n}}{B_{2,\infty}} \ .$$

The RHS is upper bounded by 1. This concludes the proof. □

A.2 Proof of Theorem 1

(i) Set

$$\begin{aligned}v_n &\leftarrow \gamma_n t_{n-1}^2 \overline{F}(\theta_n) + \frac{1}{2}\|\Delta_n - \theta_\star\|^2, \\ \chi_n &\leftarrow \left(\gamma_n t_{n-1}^2 - \gamma_{n+1} t_n (t_n - 1)\right) \overline{F}(\theta_n), \\ b_n &\leftarrow \gamma_{n+1}^2 t_n^2 \|\eta_{n+1}\|^2 - \gamma_{n+1} t_n \left\langle \overline{\Delta}_n - \theta_\star, \eta_{n+1} \right\rangle ;\end{aligned}$$

so that by Lemma 8, $v_{n+1} \leq v_n - \chi_n + b_n$. We apply Lemma 17 since $\sum_n b_n$ exists under (8). This yields (9) and

$$\lim_n \left(\gamma_{n+1} t_n^2 \overline{F}(\theta_n) + \|\Delta_n - \theta_\star\|\right) \text{ exists},$$

from which we deduce (10).

The property $\sup_n \|\Delta_n - \theta_\star\| < \infty$ yields $\sup_n \|\Delta_n\| < \infty$. By Lemma 7 and the assumption (8), we also have $\sup_n \|\overline{\Delta}_n\| < \infty$.

(ii) The proof follows from the convexity of F and the iteration of (18).

A.3 Proof of Theorem 3

Let B_k be given by Lemma 11. The stated assumptions imply that $\sum_n \sum_{k=2}^{n+1} \left(\prod_{j=k}^n \alpha_j\right) B_k$ is finite. The result follows from Lemma 11 and Lemma 17 applied with $v_n \leftarrow \|\theta_n - \theta_\star\|^2$ and $b_n \leftarrow \sum_{k=2}^{n+1} \left(\prod_{j=k}^n \alpha_j\right) B_k$.

A.4 Proof of Theorem 5

Proof of the first claim We show that the assumptions of Theorem 1 hold almost-surely, which will imply that its conclusion holds almost-surely; by Lemma 18i, $t_{n-1}^2 - t_n(t_n - 1) \geq O(n^d)$, which yields the result.

Let us prove that the assumptions hold almost-surely. By H4, there exists a constant C such that $t_n^2 \leq C n^{2d}$. Combined with H5, this yields

$$\mathbb{E}\left[\sum_n t_n^2 \|\eta_{n+1}\|^2\right] \leq 2 \sum_n t_n^2 \mathbb{E}\left[\|\epsilon_{n+1}\|^2 + \|\xi_{n+1}\|^2\right] \leq C \sum_n n^{2d} \left(\frac{C_\epsilon}{n^{2\mathsf{a}}} + \frac{C_\xi}{n^{2\mathsf{b}}}\right).$$

We write $\sum_{k=0}^n t_k \left\langle \overline{\Delta}_k - \theta_\star, \eta_{k+1} \right\rangle = \mathcal{T}_{1,n} + \mathcal{T}_{2,n}$ with $\mathcal{T}_{1,n} \stackrel{\text{def}}{=} \sum_{k=0}^n t_k \left\langle \overline{\Delta}_k - \theta_\star, \epsilon_{k+1} \right\rangle$. Under H5, $\{\mathcal{T}_{1,n}, n \geq 0\}$ is a $\mathcal{F}_n$-adapted martingale. It converges almost surely as soon as $\sum_n t_n^2 \mathbb{E}\left[\|\overline{\Delta}_n - \theta_\star\|^2 \|\epsilon_{n+1}\|^2\right] < \infty$ (see (Hall and Heyde, 1980, Theorem 2.18)): we have by H5

$$\mathbb{E}\left[\|\overline{\Delta}_n - \theta_\star\|^2 \|\epsilon_{n+1}\|^2\right] = \mathbb{E}\left[\|\overline{\Delta}_n - \theta_\star\|^2 \mathbb{E}\left[\|\epsilon_{n+1}\|^2 | \mathcal{F}_n\right]\right] \leq \frac{C_\epsilon}{n^{2\mathsf{a}}} \sup_n \mathbb{E}\left[\|\overline{\Delta}_n - \theta_\star\|^2\right].$$

Therefore, by using Proposition 12, the martingale converges almost-surely as soon as $C_\epsilon \sum_n n^{2(d-\mathsf{a})} < \infty$. The random variable $\lim_n \mathcal{T}_{2,n}$ exists a.s. if

$$\sum_n t_n \mathbb{E}\left[\|\overline{\Delta}_n - \theta_\star\| \, \|\xi_{n+1}\|\right] < \infty;$$

by applying the Cauchy-Schwarz inequality, Proposition 12 and H5, it holds true if $C_\xi \sum_n n^{d-\mathsf{b}} < \infty$.

Proof of the second claim It follows from Proposition 14.

Proof of the third claim By H3, there exists a converging subsequence $\{\theta_{\phi_n}, n \in \mathbb{N}\}$. The limiting value of this subsequence is in $\theta_\star \in \mathcal{L}$ by Corollary 2i. Hence, $\lim_n \|\theta_{\phi_n} - \theta_\star\| = 0$.

On the other hand, by Lemma 18, Proposition 14 and Proposition 15, the assumptions of Theorem 3 hold. Hence $\lim_n \|\theta_n - \theta\|$ exists for any $\theta \in \mathcal{L}$.

Combining these results yield the claim since $\lim_n \|\theta_n - \theta_\star\| = \lim_n \|\theta_{\phi_n} - \theta_\star\|$.

A.5 Technical lemmas

Lemma 16. *Assume H1. For all $\theta, \vartheta, \xi \in \Theta$ and $\gamma \in (0, 1/L]$,*

$$-2\gamma \left(F(\mathrm{Prox}_{\gamma,g}(\theta)) - F(\vartheta)\right) \geq \|\mathrm{Prox}_{\gamma,g}(\theta) - \vartheta\|^2 + 2\left\langle \mathrm{Prox}_{\gamma,g}(\theta) - \vartheta, \xi - \gamma \nabla f(\xi) - \theta \right\rangle - \|\vartheta - \xi\|^2 .$$

Proof. See (Atchade et al., 2017, Lemma 8). □

Lemma 17. *Let $\{v_n,\ n \in \mathbb{N}\}$ and $\{\chi_n,\ n \in \mathbb{N}\}$ be non-negative sequences and $\{b_n,\ n \in \mathbb{N}\}$ be such that $\sum_n b_n$ exists. If for any $n \geq 0$, $v_{n+1} \leq v_n - \chi_n + b_n$ then $\sum_n \chi_n < \infty$ and $\lim_n v_n$ exists.*

Proof. See (Atchade et al., 2017, Lemma 1). □

Lemma 18. *Assume H4. Then*

(i) $t_{n-1}^2 - t_n(t_n - 1) \geq t_n(1 - (2d)/a^d)$ and the condition (3) is satisfied.

(ii) for any $n \geq 2$,

$$\frac{t_n - 1}{t_n} \geq 1 - \left(\frac{a}{1+a}\right)^d, \qquad \text{and} \qquad t_n^2 - (t_{n-1} - 1)^2 \geq t_n$$

(iii) for any $n \geq 2$,

$$\sup_{k \geq 2} \frac{1}{t_k} \sum_{m \geq k} \prod_{n=k}^{m} \frac{t_n - 1}{t_{n+1}} < \infty .$$

Proof. Proof of (i) See (Aujol and Dossal, 2015, Lemma 2). *Proof of* (ii) The LHS follows from

$$t_n^{-1} = \left(\frac{a}{n + a - 1}\right)^d \leq \left(\frac{a}{1+a}\right)^d \text{ for any } n \geq 2 . \tag{28}$$

For the RHS, we write $t_n^2-(t_{n-1}-1)^2 = (t_n-t_{n-1}+1)(t_n+t_{n-1}-1)$. Since $t_n \geq t_{n-1}$, the first term is lower bounded by 1. By (28), the second term is lower bounded by

$$t_n\left(1-\frac{1}{t_n}+\left(\frac{n+a-2}{n+a-1}\right)^d\right) \geq t_n\left(1-\left(\frac{a}{1+a}\right)^d+\left(1-\frac{1}{n+a-1}\right)^d\right)$$
$$\geq t_n\left(1-\left(\frac{a}{1+a}\right)^d+\left(1-\frac{1}{1+a}\right)^d\right) = t_n\ .$$

Proof of (iii) See (Aujol and Dossal, 2015, Lemma 7). □

Lemma 19. *Let* $\{u_n,\ n\in\mathbb{N}\}$, $\{v_n,\ n\in\mathbb{N}\}$ *and* $\{e_n,\ n\in\mathbb{N}\}$ *be sequences satisfying* $u_n^2 \leq v_n + \sum_{k=0}^n u_k e_k$ *and* $2v_n + \sum_{k=0}^n e_k^2 \geq 0$. *Set* $\mathcal{U}(a,b) \stackrel{\text{def}}{=} b+\sqrt{a+b^2}$. *Then for any* $n \geq 0$,

$$\sup_{0\leq k\leq n}\left|u_k-\frac{e_k}{2}\right| \leq \mathcal{U}\left(v_n+\frac{1}{2}\sum_{k=0}^n e_k^2, \frac{1}{2}\sum_{k=0}^{n-1}|e_k|\right)$$

with the convention that $\sum_{k=0}^{-1}=0$.

Proof. The proof is adapted from (Schmidt et al., 2011, Lemma 1). For any $n \geq 1$,

$$\left(u_n-\frac{e_n}{2}\right)^2 \leq v_n+\frac{1}{4}e_n^2+\sum_{k=0}^{n-1}u_k e_k \leq v_n+\frac{1}{2}\sum_{k=0}^n e_k^2+\sum_{k=0}^{n-1}\left(u_k-\frac{e_k}{2}\right)e_k\ .$$

Set

$$A_n \stackrel{\text{def}}{=} v_n+\frac{1}{2}\sum_{k=0}^n e_k^2 \qquad B_n \stackrel{\text{def}}{=} \frac{1}{2}\sum_{k=0}^n |e_k| \qquad s_n \stackrel{\text{def}}{=} \sup_{0\leq k\leq n}\left|u_k-\frac{e_k}{2}\right|\ .$$

Then $s_n^2 \leq s_{n-1}^2 \vee \{A_n + s_{n-1}\,2B_{n-1}\}$. By induction (note that $s_0 \leq \sqrt{A_0}$ and $B_{-1}=0$), this yields for any $n \geq 0$,

$$0 \leq s_n \leq B_{n-1}+\left(B_{n-1}^2+A_n\right)^{1/2}\ .$$

□

References

Atchade, Y., Fort, G. and Moulines, E. (2014). On Stochastic Proximal-Gradient algorithms. Tech. rep., arXiv 1402.2365-v1.

Atchade, Y., Fort, G. and Moulines, E. (2017). On Perturbed Proximal Gradient Algorithms. *Journal of Machine Learning Research* **18** 1–33.

Aujol, J. and Dossal, C. (2015). Stability of over-relaxations for the Forward-Backward algorithm; application to FISTA. *SIAM Journal on Optimisation* **25**. (personnal communication).

Bauschke, H. H. and Combettes, P. L. (2011). *Convex analysis and monotone operator theory in Hilbert spaces*. CMS Books in Mathematics/Ouvrages de Mathématiques de la SMC, Springer, New York. With a foreword by Hédy Attouch. URL `http://dx.doi.org/10.1007/978-1-4419-9467-7`

Beck, A. and Teboulle, M. (2009). A Fast Iterative Shrinkage-Thresholding Algorithm for Linear Inverse Problems. *SIAM J. Imaging Sci.* **2** 183–202.

Chambolle, A. and Dossal, C. (2014). On the Convergence of the Iterates of "FISTA". Tech. rep., HAL-01060130v3.

Chen, L. (1978). A short note on the Conditional Borel-Cantelli Lemma. *Ann. Probab.* **6** 699–700.

Fort, G., Ollier, E. and Samson, A. (2018a). Stochastic Proximal-Gradient algorithms for penalized mixed models. *Statistics and Computing* **29** 231–253.

Fort, G., Risser, L., Atchade, Y. and Moulines, E. (2018b). Stochastic FISTA Algorithms: so fast ? In *Proceedings of the IEEE Workshop in Statistical Signal Processing*.

Hall, P. and Heyde, C. (1980). *Martingale Limit Theory and its Application*. Academic Press.

Oates, C. J., Girolami, M. and Chopin, N. (2017). Control functionals for Monte Carlo integration. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* **79** 695–718.

Schmidt, M., Le Roux, N. and Bach, F. (2011). Convergence rates of inexact proximal-gradient methods for convex optimization. In *NIPS*. See also the technical report INRIA-00618152.