STOCHASTIC FISTA ALGORITHMS: SO FAST ?

G. Fort[1]*, L. Risser*[1]*, Y. Atchadé*[2]*, E. Moulines*[3]*,*

[1] IMT, Université de Toulouse & CNRS, F-31062 Toulouse, France.
[2] Department of Statistics, Univ. of Michigan, 1085 South University Ave, Ann Arbor 48109, MI, USA.
[3] CMAP, Ecole Polytechnique, Route de Saclay,91128 Palaiseau Cedex, France.

ABSTRACT

Motivated by challenges in Computational Statistics such as Penalized Maximum Likelihood inference in statistical models with intractable likelihoods, we analyze the convergence of a stochastic perturbation of the Fast Iterative Shrinkage-Thresholding Algorithm (FISTA), when the stochastic approximation relies on a biased Monte Carlo estimation as it happens when the points are drawn from a Markov chain Monte Carlo (MCMC) sampler. We first motivate this general framework and then show a convergence result for the perturbed FISTA algorithm. We discuss the convergence rate of this algorithm and the computational cost of the Monte Carlo approximation to reach a given precision. Finally, through a numerical example, we explore new directions for a better understanding of these Proximal-Gradient based stochastic optimization algorithms.

Index Terms— Computational Statistics, Stochastic Approximation, Markov chain Monte Carlo, Proximal-Gradient algorithms, Nesterov acceleration.

1. INTRODUCTION

In various analyses, we are faced with solving:

$$\mathrm{argmin}_{\theta\in\Theta}\left(f(\theta)+g(\theta)\right)\,, \tag{1}$$

where the set Θ and the functions f, g satisfy

A1 *$g:\mathbb{R}^d\to[0,+\infty]$ is convex, not identically $+\infty$ and lower semi-continuous; $f:\mathbb{R}^d\to\mathbb{R}\cup\{+\infty\}$ is continuously differentiable on $\Theta:=\{\theta\in\mathbb{R}^d : g(\theta)+|f(\theta)|<\infty\}$ and its gradient is L-Lipschitz on Θ;*

and the gradient ∇f is numerically intractable. Motivated by situations arising in Computational Statistics (see the examples in Section 2), we consider the case when

A2 *for any $\theta\in\mathbb{R}^d$, $\nabla f(\theta)=\int_{\mathsf{X}} H(\theta,x)\,\pi_\theta(\mathrm{d}x)$ where X is a topological space endowed with its Borel σ-field, π_θ is a probability measure on X and $H:\mathbb{R}\times\mathsf{X}\to\mathbb{R}^d$ is measurable. In addition, $x\mapsto H(\theta,x)$ is π_θ-integrable for any $\theta\in\mathbb{R}^d$,*

and only an approximation of $\nabla f(\theta)$ is available, possibly a stochastic approximation and if such, possibly a biased one. In the present paper, our main contribution is to address a convergence analysis of a numerical tool to solve Eq.(1), namely a Stochastic perturbation of FISTA (see [1]), in the challenging situation when the perturbation comes from a stochastic and biased approximation of ∇f.

This work is partially supported by ANR-11-LABX-0040-CIMI within the program ANR-11-IDEX-0002-02

2. PENALIZED MAXIMUM LIKELIHOOD ESTIMATION IN MODELS WITH INTRACTABLE LIKELIHOOD

In this section, two classes of problems arising in Computational Statistics, and illustrating the question (1) in the framework A1-A2, are presented. The first situation corresponds to the computation of the Penalized Maximum Likelihood, or equivalently the Bayesian Maximum a Posteriori estimator, in latent variable models. In that case, g stands for the penalty term on parameter θ (in the Bayesian context, the prior on the parameter); while f is the normalized negative log-likelihood: for latent variable models, it is of the form (see e.g.[2])

$$f(\theta)=-\ell_N(\theta):=-\frac{1}{N}\log\int_{\mathsf{X}} p(x,\theta)\mathrm{d}\mu(x) \tag{2}$$

where for any θ, $p(\cdot,\theta)\mathrm{d}\mu$ is the complete data likelihood and the latent variables x take values in X (μ is a positive σ-finite measure, such as the Lebesgue measure when $\mathsf{X}\subseteq\mathbb{R}^p$ or the counting measure when X is countable). In (2), the dependence upon the N observations is omitted. Under regularity conditions on the model,

$$\nabla f(\theta)=-\frac{1}{N}\int_{\mathsf{X}}\partial_\theta\log p(x,\theta)\;\mathrm{d}\pi_\theta(x) \tag{3}$$

where

$$\mathrm{d}\pi_\theta(x):=\frac{p(x,\theta)\,\mathrm{d}\mu(x)}{\int_{\mathsf{X}} p(u,\theta)\,\mathrm{d}\mu(u)}=\frac{p(x,\theta)\,\mathrm{d}\mu(x)}{\exp(N\ell_N(\theta))} \tag{4}$$

is the a posteriori distribution (of the latent variables, given the observations, when the parameter is θ) which is known up to a normalizing constant. In this example, the computation of the gradient ∇f is not explicit; the gradient is an expectation with respect to a distribution known up to a normalizing constant; this integral can be approximated by a Monte Carlo sum computed from the output of an MCMC sampler (see e.g. [3, Chapter 6]), thus providing a biased stochastic approximation of the exact gradient: note indeed that if $\{X_{j,\theta}, j\geq 0\}$ is a (non stationary) ergodic Markov chain produced by an MCMC sampler with target $\mathrm{d}\pi_\theta$, then for any positive measurable function h

$$\mathbb{E}\left[\frac{1}{m}\sum_{j=1}^{m}h(X_{j,\theta})\right]-\int h\,\mathrm{d}\pi_\theta\neq 0$$

but this bias vanishes when $m\to\infty$ (see e.g. [4, Chapter 13]).

The second situation corresponds to the computation of the Penalized Maximum Likelihood estimator in a binary graphical model.

Denote by $\mathsf{Y} := (Y^{(1)}, \cdots, Y^{(N)})$ the N $\{0,1\}^p$-valued observations, modeled as independent and identically distributed under the distribution

$$\pi_\theta(\mathrm{d}x) := \frac{1}{Z_\theta} \exp\left(\sum_{i=1}^{p} \theta_i x_i + \sum_{1 \le i < j \le p} \theta_{ij} \mathbb{1}_{x_i = x_j} \right) \mu(\mathrm{d}x)$$

where $x = (x_1, \cdots, x_p) \in \{0,1\}^p$, μ is the counting measure on $\mathsf{X} := \{0,1\}^p$, and Z_θ is the normalizing constant

$$Z_\theta := \sum_{u \in \{0,1\}^p} \exp\left(\sum_{i=1}^{p} \theta_i u_i + \sum_{1 \le i < j \le p} \theta_{ij} \mathbb{1}_{u_i = u_j} \right), \quad (5)$$

which is intractable when p is large (the sum is over 2^p points). In this example, since the unknown parameter is of length $d = p(p+1)/2$ which is often far larger than the number of observations N, g is introduced as a regularization term and f is again the normalized negative log-likelihood:

$$f(\theta) = -\ell_N(\theta) := -\log Z_\theta + \sum_{i=1}^{p} \theta_i \left(\frac{1}{N} \sum_{n=1}^{N} Y_i^{(n)} \right) + \sum_{1 \le i < j \le p} \theta_{ij} \left(\frac{1}{N} \sum_{n=1}^{N} \mathbb{1}_{Y_i^{(n)} = Y_j^{(n)}} \right).$$

The gradient of f is given by

$$\partial_{\theta_i} f(\theta) := -\int_{\mathsf{X}} x_i \, \mathrm{d}\pi_\theta(x) + \frac{1}{N} \sum_{n=1}^{N} Y_i^{(n)} \quad (6)$$

$$\partial_{\theta_{ij}} f(\theta) := -\int_{\mathsf{X}} \mathbb{1}_{x_i = x_j} \, \mathrm{d}\pi_\theta(x) + \frac{1}{N} \sum_{n=1}^{N} \mathbb{1}_{Y_i^{(n)} = Y_j^{(n)}}. \quad (7)$$

Hence, ∇f is intractable since it is equal, up to an explicit quantity depending upon the observations Y, to an expectation over a finite set of cardinal 2^p; the distribution $\mathrm{d}\pi_\theta$ is known up to a normalizing constant, so these expectations can be approximated by a biased Monte Carlo sum computed from the output of an MCMC sampler targeting $\mathrm{d}\pi_\theta$.

A numerical tool to overcome the intractability of f could rely on an approximation of this function itself, combined with an optimization technique of zero-order. For example, in (2), the intractable integral could be approximated by a Monte Carlo sum with points sampled under the distribution of the missing data x; and in (5), the intractable sum could be approximated by a Monte Carlo sum with points sampled under the uniform distribution on $\{0,1\}^p$. Nevertheless, in Statistics, this strategy is known to be inefficient, since it consists in sampling points (in the first example, it consists in imputing the missing variables) under a distribution which does not take into account the statistical information (such as the observations in the first example, or the statistical model in the second example). On the contrary, as shown by (3) and (6)-(7), an optimization method based on first-order techniques necessitates a Monte Carlo approximation that is able to take into account this knowledge: in the first example, the samples are drawn under the a posteriori distribution (see (4)) and in the second example, the samples are drawn under the statistical model π_θ. This explains the success of first-order optimization techniques to overcome the numerical intractability of Maximum Likelihood estimators.

3. PERTURBED FISTA

The examples introduced in Section 2 motivate the use of gradient-based iterative methods for solving (1). Since g is not necessarily differentiable, a natural idea to generalize gradient descent procedures is to combine explicit and implicit gradients. The convexity assumption on g allows to define the proximal operator [5]: for any $\gamma > 0$, $\tau \in \mathbb{R}^d$

$$\mathrm{Prox}_{\gamma,g}(\tau) := \mathrm{argmin}_{\theta \in \Theta} \left(g(\theta) + \frac{1}{2\gamma} \|\theta - \tau\|^2 \right); \quad (8)$$

here, $\|\cdot\|$ denotes the Euclidean norm associated with the scalar product $\langle \cdot, \cdot \rangle$. Solving the optimization problem (8) is equivalent to solving an implicit gradient equation involving the sub-differential of g [6]. On the other hand, Nesterov introduced an acceleration scheme of the deterministic gradient method that is known to converge at a rate $O(1/n^2)$, where n is the number of iterations (see [7]). Combining this gradient approach for f, the proximal operator for g, and the Nesterov acceleration, yields FISTA [1]. Given two positive step-size sequences $\{\gamma_n, n \ge 1\}$ and $\{t_n, n \ge 0\}$, define the FISTA sequence $\{\theta_n, n \ge 1\}$ by: for $n \ge 1$,

$$\vartheta_n := \theta_n + \frac{t_{n-1} - 1}{t_n} (\theta_n - \theta_{n-1}) \quad (9)$$

$$\theta_{n+1} := \mathrm{Prox}_{\gamma_{n+1},g} (\vartheta_n - \gamma_{n+1} \nabla f(\vartheta_n)); \quad (10)$$

$\theta_0 \in \Theta$, $t_0 = 1$. The convergence of the sequence $\{(f+g)(\theta_n), n \ge 0\}$ to $\min(f+g)$ at the rate $O(1/n^2)$ is known (see [1, Section 4]) under convenient assumptions; the link between this algorithm and a time discretization of a second-order Ordinary Differential Equation was recently addressed (see [8], see also [9]). A natural extension of FISTA to the case when only an approximation H_{n+1} of $\nabla f(\vartheta_n)$ is available, consists in replacing (10) with

$$\theta_{n+1} := \mathrm{Prox}_{\gamma_{n+1},g} (\vartheta_n - \gamma_{n+1} H_{n+1}) \quad (11)$$

thus introducing at each iteration of the algorithm, an error $\mathcal{E}_{n+1} := \gamma_{n+1} (H_{n+1} - \nabla f(\vartheta_n))$ before applying the proximal operator. When, in the framework A2, this approximation is a Monte Carlo sum computed from the output $\{X_{jn}, j \ge 0\}$ of an MCMC sampler with target distribution $\mathrm{d}\pi_{\vartheta_n}$, the error is

$$\mathcal{E}_{n+1} = \gamma_{n+1} \left(\frac{1}{m_{n+1}} \sum_{j=1}^{m_{n+1}} H(\vartheta_n, X_{jn}) - \int H(\vartheta_n, \cdot) \, \mathrm{d}\pi_{\vartheta_n} \right);$$

m_{n+1} is the number of Monte Carlo points at iteration $n+1$.

A natural question, which we now address, is about the conditions on the perturbations $\{\mathcal{E}_n, n \ge 0\}$ which guarantee the convergence of Perturbed FISTA (P-FISTA). We provide sufficient conditions on the approximation

$$\eta_{n+1} := H_{n+1} - \nabla f(\vartheta_n)$$

and on the sequences $\{\gamma_n, n \ge 1\}$, $\{t_n, n \ge 0\}$ ensuring that P-FISTA described by (9) and (11) inherits the same limiting behavior as FISTA when the number of iterations n tends to infinity. When $t_n = 1$ for any $n \ge 1$, P-FISTA is a Perturbed Proximal-Gradient (P-PG) algorithm [10]: a convergence analysis with an emphasis to the case of a biased Monte Carlo approximation of the gradient can be found in [11]; see also [12] for the case ∇f is of the form

$$\nabla f(\theta) = \phi(\theta) + \left\langle \overline{S}(\theta), \psi(\theta) \right\rangle, \quad \overline{S}(\theta) := \int S(x) \, \mathrm{d}\pi_\theta(x), \quad (12)$$

as it occurs in (6)-(7) and in (3) when $p(\cdot,\theta)$ is from a curved exponential family; and for a discussion on the link between this algorithm and the Stochastic Expectation-Maximization algorithms. Theorem 1 establishes the convergence of the sequence $\{(f+g)(\theta_n), n \geq 0\}$ to $\min(f+g)$ and provides an explicit control of this convergence. These results are established under the following assumptions. Set

$$z_n := \theta_n + t_n \left(\mathrm{Prox}_{\gamma_{n+1},g}(\vartheta_n - \gamma_{n+1}\nabla f(\vartheta_n)) - \theta_n\right).$$

A3 *a) The function f is convex and the set $\mathcal{L} := \mathrm{argmin}_\Theta(f+g)$ is non empty.*

b) $t_0 = 1$, and for any $n \geq 1 : t_n \geq 1$ and $\gamma_n \in]0, 1/L]$ where L is given by A1. Furthermore, $\tau_n := \gamma_n t_{n-1}^2 - \gamma_{n+1} t_n (t_n - 1) \geq 0$.

c) The series $\sum_n \gamma_{n+1} t_n \langle z_n - \theta_\star, \eta_{n+1}\rangle$ exists for some $\theta_\star \in \mathcal{L}$.

Theorem 1 *Let $\{\theta_n, n \geq 0\}$ be the P-FISTA sequence. Set $F(\theta) := (f+g)(\theta) - \min(f+g)$. Assume A1 and A3a-c). Then $\lim_n \gamma_{n+1} t_n^2 F(\theta_n)$ exists and for $\theta_\star$ given by A3c),*

$$\gamma_{n+1} t_n^2 F(\theta_{n+1}) + \frac{1}{2}\|z_n - \theta_\star\|^2 + \sum_{k=1}^n \tau_k F(\theta_k) \leq B_n$$

where τ_k is given by A3b) and

$$B_n := \gamma_1 F(\theta_1) + \frac{1}{2}\|\theta_1 - \theta_\star\|^2 + \sum_{k=1}^n \gamma_{k+1}^2 t_k^2 \|\eta_{k+1}\|^2 - \sum_{k=1}^n \gamma_{k+1} t_k \langle z_k - \theta_\star, \eta_{k+1}\rangle$$

is uniformly bounded.

The proof of this theorem is given in the technical report [13, Theorem 13]. It relies on the key property (see [13, Lemma 23]):

$$\tau_k F(\theta_k) + \gamma_{k+1} t_k^2 F(\theta_{k+1}) + \frac{1}{2}\|\tilde{z}_{k+1} - \theta_\star\|^2 \leq \gamma_k t_{k-1}^2 F(\theta_k) + \frac{1}{2}\|\tilde{z}_k - \theta_\star\|^2 - \gamma_{k+1} t_k \langle \tilde{z}_{k+1} - \theta_\star, \eta_{k+1}\rangle,$$

for any $k \geq 1$, where $\tilde{z}_k := \theta_k + t_k(\theta_{k+1} - \theta_k)$. This theorem extends the results in [1] to the case $\eta_n \neq 0$. It generalizes [8, Theorems 5.1. and 5.2.] which address the case when $t_n = 1 + n/(\alpha - 1)$ for some $\alpha \geq 3$ and $\gamma_k = \gamma \in]0, 1/L]$ for any $k \geq 1$. Theorem 1 covers both deterministic and stochastic perturbations. In the stochastic case, this general statement has two major consequences: first, when the assumption A3c) holds almost-surely, then the conclusion of the theorem holds almost-surely; second, the explicit control of $F(\theta_n)$ through the expression of B_n can be used to derive almost-sure controls of the convergence of $(f+g)(\theta_n)$ to $\min(f+g)$, but also L^p-controls, controls with high probability, etc.

We now apply this theorem in the context A2 when H_{n+1} is a Monte Carlo sum computed with m_{n+1} points $\{X_{jn}, j \leq m_{n+1}\}$ sampled by an MCMC sampler with target $\mathrm{d}\pi_{\vartheta_n}$. When $\{\gamma_n, n \geq 1\}$ is non-increasing, the condition A3b) is satisfied with $t_n = O(n)$ (choose for example $t_n = 1 + n/2$). Checking the condition A3c) is more or less technical, depending on the Monte Carlo strategy: does $m_n \to +\infty$ or is it constant over iterations? The second situation is clearly the most technical (see e.g. the mathematical derivations for the study of P-PG algorithms in [11, Section 3.1.]) and in this paper, we only address the case when $\lim_n m_n = +\infty$: it means that the computational cost of the Monte Carlo approximation increases along iterations, but this yields a vanishing bias of the Monte Carlo error η_{n+1} when $n \to \infty$. Under suitable conditions on the MCMC sampler (see [11, Assumption H5]) and assuming Θ is bounded, we have (see [11, Proposition 5])

$$\sup_n m_{n+1}^2 \, \mathbb{E}\left[\|\mathbb{E}\left[\eta_{n+1}|\mathcal{F}_n\right]\|^2\right] < \infty, \tag{13}$$

$$m_{n+1} \, \mathbb{E}\left[\|\eta_{n+1} - \mathbb{E}\left[\eta_{n+1}|\mathcal{F}_n\right]\|^2 |\mathcal{F}_n\right] \leq C_n \text{ a.s.}, \tag{14}$$

where $\mathcal{F}_n$ is the σ-field associated with the past of the algorithms $\{X_{jk}, 1 \leq j \leq m_{k+1}, k \leq n-1\}$ and C_n is a random variable such that $\sup_n \mathbb{E}[C_n^p] < \infty$ for some $p > 1$. Using standard tools for the convergence of martingales (see e.g. [14]) and since a series with positive general term converges almost-surely as soon as its expectation is finite, we prove that A3c) holds if (see [13, Corollary 14])

$$\sum_n \gamma_{n+1} t_n m_{n+1}^{-1} + \sum_n \gamma_{n+1}^2 t_n^2 m_{n+1}^{-1} < \infty. \tag{15}$$

Let us detail these conditions in the case $t_n = O(n)$, $\gamma_n \sim \gamma_\star n^{-\mathsf{a}}$ for some $a \in [0, 2[$ and $m_n \sim m_\star (\ln n)^{\mathsf{b}} n^{\mathsf{c}}$; the conditions $\lim_n \gamma_{n+1} t_n^2 = +\infty$ and (15) are verified by choosing

$$\text{either } (\mathsf{a} \in [0,1], \ \mathsf{c} = 3 - 2\mathsf{a}) \text{ or } (\mathsf{a} \in [1,2[, \ \mathsf{c} = 2 - \mathsf{a}) \tag{16}$$

and $\mathsf{b} > 1$. In both cases, we have

$$\lim_n n^{2-\mathsf{a}} F(\theta_n) \qquad \text{exists a.s.} \tag{17}$$

and by using the expression of B_n in Theorem 1, we also have

$$\sup_n n^{2-\mathsf{a}} \, \mathbb{E}\left[F(\theta_n)\right] < \infty. \tag{18}$$

From Theorem 1 and (15), we also obtain

$$\sup_n \sum_{k=1}^n k F(\theta_k) < \infty \text{ a.s.}, \qquad \sup_n \sum_{k=1}^n k \, \mathbb{E}\left[F(\theta_k)\right] < \infty. \tag{19}$$

Eqs. (17) and (18) show that the best convergence rate of the sequence $\{(f+g)(\theta_n), n \geq 0\}$ to $\min(f+g)$ is n^{-2}, for the almost-sure convergence and the convergence in expectation as well. This rate is similar to the rate of FISTA (see [1] and [8], itself known to be optimal) and in that sense, it is optimal. Nevertheless, P-FISTA reaches this rate when $\mathsf{a} = 0$, which means that the number of Monte Carlo points m_n increases at the rate $O(n^3)$ up to logarithmic terms (we have indeed by (16): $\mathsf{c} = 3 - 2\mathsf{a} = 3$). Therefore, since the number of iterations to reach a given precision $\delta > 0$ is $O(\delta^{-1/2})$, the Monte Carlo computational cost increases like $O\left(\sum_{k=1}^{\delta^{-1/2}} k^3\right)$ that is like $O(\delta^{-2})$. The Monte Carlo computational cost is unchanged if we choose $\mathsf{a} \in]0, 1[$ and $\mathsf{c} = 3 - 2\mathsf{a}$: the rate of convergence after n iterations is $O(n^{\mathsf{a}-2})$, the number of iterations to reach a precision δ is $O(\delta^{1/(\mathsf{a}-2)})$, the computational cost increases like $O(\delta^{(4-2\mathsf{a})/(\mathsf{a}-2)}) = O(\delta^{-2})$.

In [11, Sections 3.1. and 3.2.], we studied the rate of convergence for the P-PG algorithm (which corresponds to P-FISTA applied with $t_n = 1$ for any n): we proved that when $\gamma_n = \gamma$ and $m_n = O(n)$, the rate of convergence (in L^q, for $q \geq 1$) of $F(\bar{\theta}_n)$ to zero is $O(1/n)$ - up to logarithmic terms - where $\bar{\theta}_n := n^{-1}\sum_{k=1}^n \theta_k$ is the averaged value of the estimators (note that this averaging strategy is a post-processing of the output of P-PG). It

means that a precision $\delta > 0$ is reached after $O(1/\delta)$ iterations, thus involving $O(\delta^{-2})$ Monte Carlo draws.

As a conclusion, P-FISTA applied with convenient design parameters (the sequences $\{\gamma_n, n \geq 1\}$, $\{t_n, n \geq 0\}$ and the Monte Carlo sample size $\{m_n, n \geq 1\}$) reaches the same rate of convergence as FISTA after n iterations. This rate is far better than the rate of the PG algorithm. Nevertheless, when taking into account the Monte Carlo computational cost: combining a P-PG algorithm with an averaging strategy or applying P-FISTA, is equivalent; in both cases, $O(\delta^{-2})$ Monte Carlo samples are required in order to reach a given precision δ.

(19) extends the results of [8] to general sequences $\{t_n, n \geq 0\}$ and $\{\gamma_n, n \geq 0\}$ and to the case $\{\eta_n, n \geq 1\}$ is a biased Stochastic error. It also extends the results of [15, Theorem 2] to a non-constant sequence $\{\gamma_n, n \geq 0\}$ and to the case $\eta_n \neq 0$. When $t_n = O(n)$ and $\gamma_n \sim \gamma_\star n^{-\mathsf{a}}$ for some $a \in [0, 2[$ we have $\tau_n = O(n^{1-\mathsf{a}})$: the rate is maximal by choosing, here again, $\mathsf{a} = 0$ i.e. a constant step-size sequence $\{\gamma_n, n \geq 1\}$. Since the upper bound B_n is uniformly bounded in n, Theorem 1 implies that $\sum_n nF(\theta_n) < \infty$.

We decided to give a strong emphasis on Theorem 1 applied to the case the approximation H_{n+1} is stochastic and biased (see the quantity $\mathbb{E}[\eta_{n+1}|\mathcal{F}_n]$ in (13) which is allowed to be non null) since our work is motivated by the applications described in Section 2. Nevertheless, all our results apply to the simpler case when the Monte Carlo approximation is computed from independent and identically distributed (i.i.d.) points. This i.i.d. context covers many applications in Machine Learning such as problems called Large scale convex optimization (see e.g. [16]) or problems related to on-line learning; in these contexts, m_n stands resp. for the number of component functions in the Monte Carlo approximation, and for the batch size in the online data processing.

4. BEYOND THE THEORY THROUGH A NUMERICAL APPLICATION

We conclude this paper by a numerical analysis of the behavior of some algorithms that are not fully understood yet, and for which a theoretical study is in progress: we explore different strategies for the sequence $\{t_n, n \geq 0\}$ and for a Monte Carlo approximation of the gradient. As a numerical example, we consider the optimization of a likelihood in a binary graphical model (see Section 2), in the case the penalty term is

$$g(\theta) = \lambda \sum_{1 \leq i < j \leq p} |\theta_{ij}| + \mu \sum_{i=1}^{p} \theta_i^2,$$

that is, a sparsity constraint on the off-diagonal parameters, and a quadratic penalty for the diagonal ones. N independent data are sampled (set as the last values of a long run of N independent Markov chains with stationary distribution $\pi_{\theta_{\text{true}}}$; the "true" parameter θ_{true} was chosen as a sparse upper triangular matrix, 195 off-diagonal entries are non zero. In the numerical illustration, $\lambda = 0.5\sqrt{\log(p)/N}$, $\mu = 0.5$, $N = 250$, $p = 100$ and we use the Wolff clustering algorithm [17] to sample a Markov chain with target π_θ.

Four algorithms are compared: `Alg1` is a P-PG algorithm (that is P-FISTA with $t_n = 1$ for any n) with $\gamma_n = O(1/\sqrt{n})$ and $m_n = O(\sqrt{n})$. `Alg2` to `Alg4` are P-FISTA's, all of them with $\gamma_n = O(1)$ and $m_n = O(n^3)$ but they differ through the choice of the sequence $\{t_n, n \geq 0\}$ which is respectively of the form $O(n)$, $O(\sqrt{n})$ and $O(n^\epsilon)$ for some $\epsilon << 1$. Since in our example, ∇f is of the form (12), then $H_{n+1} = \phi(\theta_n) + \langle S_{n+1}, \psi(\theta_n)\rangle$ where

$$S_{n+1} := m_{n+1}^{-1} \sum_{j=1}^{m_{n+1}} S(X_{jn}), \tag{20}$$

for `Alg1` to `Alg4`. `Alg5` looks like `Alg1` (it is a P-PG) except that S_{n+1} is computed iteratively along the iterations, so that it uses all the past samples (see [12])

$$S_{n+1} := (1 - \delta_{n+1})S_n + \delta_{n+1} m_{n+1}^{-1} \sum_{j=1}^{m_{n+1}} S(X_{jn}); \tag{21}$$

in the numerical example, we choose $\delta_n = O(n^{-0.9})$ and $m_n = O(1)$. Each algorithm is run along $n_{\max} = 2000$ iterations, and these runs are repeated 100 times; m_n is defined in such a way that all the algorithms used roughly the same total number of Monte Carlo points after $n_{\max}$ iterations. For each of the five algorithms, we display on Figure 1[left] the boxplot (over the 100 runs) of the number of non-null off-diagonal elements at iteration n, when $n \in \{50, 500, 1000, 1500, 2000\}$ iterations. We also display on Figure 1[right] how often, over the 100 runs, each entry of $\theta_{n_{\max}}$ is non null: a white vertical line is 0%, a black one is 100%; the entries are sorted so that the first p elements are the diagonal entries (which are not affected by the L_1 penalty, thus explaining p successive black horizontal lines for all the algorithms). The first row on Figure 1[right] shows, as a reference, the vector θ_{true}.

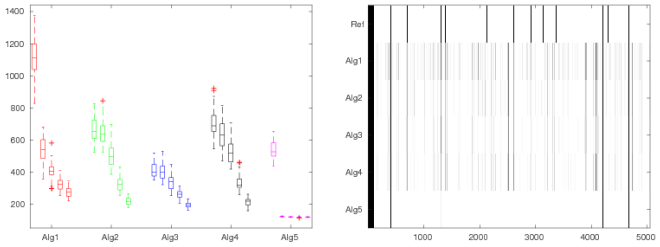

Fig. 1. For different algorithms, [left] the number of non zero components of θ_n when $n \in \{50, 500, 1000, 1500, 2000\}$; [right] after $n_{\max}$ iterations, probability to be non null for each component $\theta_{ij}, 1 \leq i \leq j \leq p$.

The plot on the left illustrates that `Alg2` converges more rapidly than `Alg1`. It also shows that among the P-FISTA strategies, there may have a gain to use a sequence t_n that tends to infinity at a slower rate than $O(n)$: a rigorous theoretical analysis of these strategies for Stochastic perturbation of FISTA is an open question (see [18] for deterministic perturbations).

Comparing `Alg1` to `Alg5` shows that a kind of smoothed gradient as computed by (21) improves drastically the rate of convergence: this idea was studied in [12] for P-PG; its use in P-FISTA is also an open question.

The plot on the right illustrates again that the convergence of `Alg1` to `Alg4` is slower that the one of `Alg5`. Note that the non-null components at convergence of `Alg5` are strongly related to those of θ_{true}: how many components are equal depend of course of the choice of λ, a question which is out of the scope of this paper.

5. REFERENCES

[1] A. Beck and M. Teboulle, "A fast iterative shrinkage-thresholding algorithm for linear inverse problems," *SIAM J. Imaging Sci.*, vol. 2, no. 1, pp. 183–202, 2009.

[2] B. S. Everitt, *An introduction to latent variable models*, Monographs on Statistics and Applied Probability. Chapman & Hall, London; distributed by Methuen, Inc., New York, 1984.

[3] O. Cappé, E. Moulines, and T. Rydén, *Inference in hidden Markov models*, Springer Series in Statistics. Springer, New York, 2005.

[4] S. Meyn and R. L. Tweedie, *Markov chains and stochastic stability*, Cambridge University Press, Cambridge, second edition, 2009, With a prologue by Peter W. Glynn.

[5] J.J. Moreau, "Fonctions convexes duales et points proximaux dans un espace hilbertien," *C. R. Acad. Sci. Paris*, vol. 255, pp. 2897–2899, 1962.

[6] R.T. Rockafellar, "Monotone operators and the proximal point algorithm," *SIAM J. Control Optimization*, vol. 14, no. 5, pp. 877–898, 1976.

[7] Y. E. Nesterov, "A method for solving the convex programming problem with convergence rate $O(1/k^2)$," *Dokl. Akad. Nauk SSSR*, vol. 269, no. 3, pp. 543–547, 1983.

[8] H. Attouch and Z. Chbani, "Fast inertial dynamics and FISTA algorithms in convex optimization. Perturbation aspects," Tech. Rep., arXiv 1507.01367, 2015.

[9] A. Cabot, H. Engler, and S. Gadat, "On the long time behavior of second order differential equations with asymptotically small dissipation," *Trans. Amer. Math. Soc.*, vol. 361, no. 11, pp. 5983–6017, 2009.

[10] P. L. Combettes and J.C. Pesquet, "Proximal splitting methods in signal processing," in *Fixed-point algorithms for inverse problems in science and engineering*, vol. 49 of *Springer Optim. Appl.*, pp. 185–212. Springer, New York, 2011.

[11] Y. F. Atchadé, G. Fort, and E. Moulines, "On perturbed proximal gradient algorithms," *J. Mach. Learn. Res.*, vol. 18, pp. Paper No. 10, 33, 2017.

[12] G. Fort, E. Ollier, and A. Samson, "Stochastic Proximal Gradient for Penalized Mixed Models," *accepted for publication in Statistics and Computing (arXiv:1704.08891)*, 2018.

[13] Y. Atchadé, G. Fort, and E. Moulines, "On Stochastic Proximal Gradient Algorithms," Tech. Rep., arXiv:1402.2365v2, 2015.

[14] P. Hall and C. C. Heyde, *Martingale limit theory and its application*, Academic Press, Inc. [Harcourt Brace Jovanovich, Publishers], New York-London, 1980, Probability and Mathematical Statistics.

[15] A. Chambolle and C. Dossal, "On the convergence of the iterates of FISTA," *Journal of Optimization Theory and Applications*, vol. 166, no. 3, 2015.

[16] D. Bertsekas, *Optimization for Machine Learning*, pp. 85–119, MIT Press, Cambridge, 2012.

[17] U. Wolff, "Collective Monte Carlo updating for spin systems," *Phys. Rev. Lett.*, vol. 62, pp. 361–364, 1989.

[18] J.-F. Aujol and C. Dossal, "Stability of over-relaxations for the forward-backward algorithm, application to FISTA," *SIAM J. Optim.*, vol. 25, no. 4, pp. 2408–2433, 2015.