The perturbed prox-preconditioned spider algorithm: non-asymptotic convergence bounds

Gersende Fort, E Moulines

HAL Id: hal-03183775

https://hal.science/hal-03183775v2

Submitted on 24 May 2021

THE PERTURBED PROX-PRECONDITIONED SPIDER ALGORITHM: NON-ASYMPTOTIC CONVERGENCE BOUNDS

G. Fort[1], *E. Moulines*[2],

[1] IMT, Université de Toulouse & CNRS, F-31062 Toulouse, France.
[2] CMAP, Ecole Polytechnique, Route de Saclay, 91128 Palaiseau Cedex, France.

ABSTRACT

A novel algorithm named `Perturbed Prox-Preconditioned SPIDER (3P-SPIDER)` is introduced. It is a stochastic variance-reduced proximal-gradient type algorithm built on `Stochastic Path Integral Differential EstimatoR` (SPIDER), an algorithm known to achieve near-optimal first-order oracle inequality for nonconvex and nonsmooth optimization. Compared to the vanilla prox-SPIDER, `3P-SPIDER` uses preconditioned gradient estimators. Preconditioning can either be applied "explicitly" to a gradient estimator or be introduced "implicitly" as in applications to the EM algorithm. `3P-SPIDER` also assumes that the preconditioned gradients may (possibly) be not known in closed analytical form and therefore must be approximated which adds an additional degree of perturbation. Studying the convergence in expectation, we show that `3P-SPIDER` achieves a near-optimal oracle inequality $O(n^{1/2}/\epsilon)$ where n is the number of observations and ϵ the target precision even when the gradient is estimated by Monte Carlo methods. We illustrate the algorithm on an application to the minimization of a penalized empirical loss.

Index Terms— Statistical Learning, Large Scale Learning, Variance reduced Stochastic gradient, Finite sum optimization, Control Variates.

1. INTRODUCTION

Consider the following composite, nonconvex, and possibly nonsmooth optimization problem

$$\text{Argmin}_{s \in \mathcal{S}} \{W(s) + g(s)\}, \qquad (1)$$

where $\mathcal{S}$ is a closed convex subset of $\mathbb{R}^q$, $W : \mathcal{V} \to \mathbb{R}$ is a smooth function defined on a neighborhood $\mathcal{V}$ of $\mathcal{S}$ and $g : \mathcal{S} \to (-\infty, +\infty]$ is a proper lower semi-continuous convex function (with an easy to compute proximal term). This paper addresses the case when W has a finite-sum structure

$$W(s) = \frac{1}{n} \sum_{i=1}^{n} W_i(s), \qquad (2)$$

and the optimization problem (1) is solved by a *preconditioned-gradient* based algorithm. Optimization problems (1)-(2) often arise in machine learning. In such case, n is the number of examples which is typically very large, W_i is the loss function associated to example $\#i$, and g is a non-smooth regularization term, *e.g.* ℓ_1

norm, Elastic net, etc. The preconditioning setting may naturally arise "implicitly", for example in the stochastic finite-sum version of the Expectation-Maximization (EM) algorithm in the exponential family, which was the main motivation for this work; see Section 4 and the companion paper [1].

Minimization problems (1) and (2) cover a broad range of applications in machine learning, statistics, and signal processing; see [2]. State-of-the art methods to solve these problems rely on stochastic optimization approaches [3, 4]. In the nonconvex case, while numerical algorithms for solving the noncomposite setting (*i.e.* $g = 0$), are well-developed and have received significant attention [5, 6], methods for composite optimization remain scarce [7, 8]. The authors in [7] proposes and studies a non-composite finite-sum problem using `SVRG` estimator from [9]. This method is extended to the composite setting by applying the proximal operator of g as in the proximal-gradient scheme (see [10, 11, 12] for literature review on the proximal-gradient algorithm). This technique is based on gradients and does not use preconditioning. This scheme has been later improved with `SPIDER`, where the gradient control variate is sequentially updated to improve the estimation accuracy: `SPIDER` is known to achieve near optimal oracle complexity in nonconvex optimization [13, 14, 8].

This paper analyzes the `3P-SPIDER` algorithm designed to solve

$$s \in \mathcal{S}: \qquad 0 \in \nabla W(s) + \partial g(s), \qquad (3)$$

by combining *(i)* a variance-reduced preconditioned-gradient algorithm designed for the finite sum setting, and *(ii)* a proximal step to take into account the (non smooth) regularizer g. Furthermore `3P-SPIDER` covers the case when the preconditioned gradient - of the form $n^{-1} \sum_{i=1}^{n} \mathsf{h}_i(s)$, see A2 below - is not computable in a closed-form and is approximated: both the cases of an approximation of the full sum over n terms by a sum over a random subsample, and an approximation of the functions h_i's are considered. `3P-SPIDER` was introduced in [1] for a specific application to large scale learning solved by a Expectation Maximization-based optimization method. The main contribution of this paper is to provide explicit control of the convergence in expectation of `3P-SPIDER` and deduce complexity bounds in terms of the design parameters of this algorithm. A comparison to the state of the art gradient-based methods in terms of complexity bounds, is also provided in the case the quantities h_i are expectations and are approximated by a Monte Carlo integration: it is established that the number of Monte Carlo samples can be chosen so that `3P-SPIDER` reaches the oracle complexity bounds corresponding to the case where the $\mathsf{h}_i(s)$'s are known in closed form.

Notations. $\mathbb{R}_{+}^{\star}$ and $\mathbb{N}^{\star}$ denote respectively (resp.) the positive real line and the positive integers. For $n \in \mathbb{N}^{\star}$, set $[n]^{\star} \stackrel{\text{def}}{=} \{1, \cdots, n\}$ and $[n] \stackrel{\text{def}}{=} \{0, \cdots, n\}$. For $x \in \mathbb{R}$, $\lceil x \rceil$ is the near-

est integer greater than or equal to x. Vectors are column-vectors; for a, b in $\mathbb{R}^\ell$, $\langle a, b \rangle$ denotes the Euclidean scalar product, and $\|a\|$ the associated norm. For a matrix A, we denote by A^T and A^{-1} resp. its transpose and its inverse. I_d is the $d \times d$ identity matrix. The random variables are defined on a probability space $(\Omega, \mathcal{A}, \mathbb{P})$; $\mathbb{E}$ denotes the associated expectation. For random variables U and a sigma-field $\mathcal{F}$, $\mathbb{E}[U|\mathcal{F}]$ is the conditional expectation of U given $\mathcal{F}$. For a smooth function f, ∇f is the gradient of f. For a proper lower semi-continuous convex function g and x in its (assumed) non-empty domain, $\partial g(x)$ is the subdifferential of g at x.

2. THE 3P-SPIDER ALGORITHM

The optimization problem at hand is the problem (3) in the case when

A1 $\mathcal{S}$ is a closed convex subset of $\mathbb{R}^q$.
$W : \mathcal{V} \to \mathbb{R}$ is a continuously differentiable function on $\mathcal{V}$, an open neighborhood of $\mathcal{S}$. Its gradient ∇W is globally Lipschitz-continuous on $\mathcal{S}$ with Lipschitz constant $L_{\dot{W}}$.
$g : \mathcal{S} \to (-\infty, +\infty]$ is a proper lower semi-continuous convex function.

We consider a gradient approach for solving (3) and allow the use of a preconditioning matrix $B(s)$ which may depend on the current value of the parameter s. We assume that

A2 For any $s \in \mathcal{S}$, $B(s)$ is a $q \times q$ positive definite matrix and there exist $0 < v_{\min} \leq v_{\max}$ such that for any $s \in \mathcal{S}$, the spectrum of $B(s)$ is in $[v_{\min}, v_{\max}]$.
For all $i \in [n]^\star$, there exists a globally Lipschitz function $h_i : \mathcal{S} \to \mathbb{R}^q$, with constant L_i, such that

$$-B^{-1}(s) \nabla W(s) = \frac{1}{n} \sum_{i=1}^n h_i(s) .$$

We introduce a weighted proximal operator: for a $q \times q$ positive definite matrix B, define for any $s \in \mathbb{R}^q$ and $\gamma > 0$,

$$\mathrm{Prox}_{B, \gamma g}(s) \overset{\mathrm{def}}{=} \mathrm{Argmin}_{s' \in \mathcal{S}} \left\{ \gamma g(s') + \frac{1}{2}(s' - s)^T B(s' - s) \right\} .$$

Set $h(s) \overset{\mathrm{def}}{=} n^{-1} \sum_{i=1}^n h_i(s)$. Under A1 and A2, for any $s, s' \in \mathcal{S}$ and $\gamma > 0$, $\mathrm{Prox}_{B(s), \gamma g}(s')$ exists and is unique and, since $s' = \mathrm{Prox}_{B(s), \gamma g}(s + \gamma h(s))$ if and only if $0 \in \partial g(s) + B(s)(s' - s - \gamma h(s))$, we obtain

$$\begin{aligned} \{s \in \mathcal{S} : \mathrm{Prox}_{B(s), \gamma g}(s + \gamma h(s)) = s\} \\ = \{s \in \mathcal{S} : 0 \in \nabla W(s) + \partial g(s)\} . \quad (4) \end{aligned}$$

This property implies that the solutions of (3) are the roots of the function $s \mapsto \mathrm{Prox}_{B(s), \gamma g}(s + \gamma h(s)) - s$ restricted to $\mathcal{S}$, whatever $\gamma > 0$.

For any minibatch $\mathcal{B}$ of size b, sampled at random from $[n^\star]$ - with or without replacement, we have (see e.g. [15, Lemma 4])

$$h(s) = \frac{1}{b} \mathbb{E} \left[\sum_{i \in \mathcal{B}} h_i(s) \right] ,$$

thus implying that in the finite-sum setting, the preconditioned gradient $-B^{-1}(s) \nabla W(s)$ can be approximated by a sum with b terms where the indexes of summation i are sampled randomly:

$b^{-1} \sum_{i \in \mathcal{B}} h_i(s)$. Therefore, a natural extension of the Proximal-Gradient algorithm to the finite-sum setting would define a sequence $\{\widehat{S}_k, k \geq 0\}$ by

$$\widehat{S}_{k+1} = \mathrm{Prox}_{B(\widehat{S}_k), \gamma_{k+1} g} \left(\widehat{S}_k + \frac{\gamma_{k+1}}{b} \sum_{i \in \mathcal{B}_{k+1}} h_i(\widehat{S}_k) \right)$$

where $\{\gamma_k, k \geq 0\}$ is a positive deterministic sequence and $\{\mathcal{B}_{k+1}, k \geq 0\}$ is a sequence of minibatches of size b sampled at random from $[n]^\star$. 3P-SPIDER reduces the variance of this stochastic perturbation by the construction of a control variate, which is defined as an approximation of $h(\widehat{S}_k)$ correlated with the random variable $b^{-1} \sum_{i \in \mathcal{B}_{k+1}} h_i(\widehat{S}_k)$. This control variate is refreshed at each so-called *outer loop*, indexed by t in Algorithm 1; and then evolves along the *inner loops*, indexed by k. Finally, 3P SPIDER allows approximations on the computation of $h_i(\widehat{S}_{t,k})$, approximations denoted by $\widehat{h}_i^{t,k}$.

The algorithm is given in Algorithm 1. At the start of each outer loop $\#t$, the control variate $S_{t,0}$ is initialized (see Lines 2 and 10) in order to approximate $h(\widehat{S}_{t,-1})$; a natural idea is to choose $\mathcal{E}_t = 0$. Nevertheless, the computational cost is important since it involves a sum over n terms, and this full sum can be substituted with a sum having $b' \ll n$ indices defined by a minibatch $\mathcal{B}'_{t,0}$ sampled at random from $[n]^\star$; in that case, $\mathcal{E}_t \neq 0$. The control variate is modified at each inner loop $\#k$ (see Line 6): since $S_{t,0} \approx h(\widehat{S}_{t,-1})$, we have $S_{t,k+1} \approx h(\widehat{S}_{t,k})$ upon noting that $S_{t,k+1} - S_{t,k} \approx h(\widehat{S}_{t,k}) - h(\widehat{S}_{t,k-1})$. The key property is the choice of the same minibatch $\mathcal{B}_{t,k+1}$ when approximating $h(\widehat{S}_{t,k})$ and $h(\widehat{S}_{t,k-1})$: the correlation of these quantities is the essence of the *control variate* mechanism.

Data: $k_{\mathrm{out}}, k_{\mathrm{in}} \in \mathbb{N}^\star$; $\widehat{S}_{\mathrm{init}} \in \mathcal{S}$; $\gamma_{t,0} \geq 0, \gamma_{t,k} > 0$ for $t \in [k_{\mathrm{out}}]^\star, k \in [k_{\mathrm{in}}]^\star$.
Result: The 3P-SPIDER sequence
$\quad \{\widehat{S}_{t,k}, t \in [k_{\mathrm{in}}]^\star, k \in [k_{\mathrm{in}}]\}$
1 $\widehat{S}_{1,0} = \widehat{S}_{1,-1} = \widehat{S}_{\mathrm{init}}$;
2 $S_{1,0} = n^{-1} \sum_{i=1}^n \widehat{h}_i^{1,-1} + \mathcal{E}_1$;
3 **for** $t = 1, \cdots, k_{\mathrm{out}}$ **do**
4 $\quad$ **for** $k = 0, \ldots, k_{\mathrm{in}} - 1$ **do**
5 $\qquad$ Sample a mini batch $\mathcal{B}_{t,k+1}$ of size b in $[n]^\star$;
6 $\qquad$ $S_{t,k+1} = S_{t,k} + b^{-1} \sum_{i \in \mathcal{B}_{t,k+1}} \left(\widehat{h}_i^{t,k} - \widehat{h}_i^{t,k-1} \right)$;
7 $\qquad$ $\widehat{S}_{t,k+1/2} = \widehat{S}_{t,k} + \gamma_{t,k+1} S_{t,k+1}$;
8 $\qquad$ $\widehat{S}_{t,k+1} = \mathrm{Prox}_{B(\widehat{S}_{t,k}), \gamma_{t,k+1} g} \left(\widehat{S}_{t,k+1/2} \right)$;
9 $\quad$ $\widehat{S}_{t+1,-1} = \widehat{S}_{t,k_{\mathrm{in}}}$;
10 $\quad$ $S_{t+1,0} = n^{-1} \sum_{i=1}^n \widehat{h}_i^{t+1,-1} + \mathcal{E}_{t+1}$;
11 $\quad$ $\widehat{S}_{t+1,-1/2} = \widehat{S}_{t+1,-1} + \gamma_{t+1,0} S_{t+1,0}$;
12 $\quad$ $\widehat{S}_{t+1,0} = \mathrm{Prox}_{B(\widehat{S}_{t+1,-1}), \gamma_{t+1,0}}(\widehat{S}_{t+1,-1/2})$

Algorithm 1: The Perturbed Prox-Preconditioned SPIDER (3P-SPIDER) algorithm.

The SPIDER algorithm ([13, 14, 8]) corresponds to the case $g = 0$, $\mathcal{S} = \mathbb{R}^q$, $B(s) = I_q$ and $\widehat{h}_i^{t,k} = h_i(\widehat{S}_{t,k})$ for any $i \in [n]^\star$, $t \in [k_{\mathrm{out}}]^\star$ and $k \in [k_{\mathrm{in}} - 1]$. In the case $B(s) = I_q$, 3P-SPIDER is a perturbed proximal-gradient algorithm (see e.g. [16]); the con-

vergence analysis below addresses the non convex case (W is not assumed to be convex). In the case $g = 0$ and $\mathcal{S} = \mathbb{R}^q$, 3P-SPIDER is a Stochastic Approximation algorithm designed to find the roots of the preconditioned gradient $s \mapsto \mathsf{h}(s) = -B^{-1}(s)\nabla \mathrm{W}(s)$. Applied with $g = 0$ or $g = \chi_{\mathcal{K}}$ - the characteristic function of a closed convex set $\mathcal{K}$, 3P-SPIDER is a variance reduced incremental EM algorithm (see [17, 1], see also section 4 for an application to the minimization of a penalized empirical loss).

3. CONVERGENCE IN EXPECTATION AND COMPLEXITY BOUNDS

For ease of exposition (see [18] for the general case), it is assumed hereafter that

A3 $\gamma_{t+1,0} = 0$ and $\mathcal{E}_{t+1} = 0$ for any $t \in [k_{\mathrm{out}}]^\star$.
The intractable functions $s \mapsto \mathsf{h}_i(s)$ are defined as an integral with respect to (w.r.t.) a distribution $\pi_{i,s}$

$$\mathsf{h}_i(s) \stackrel{\text{def}}{=} \int_{\mathsf{Z}} \mathcal{H}_i(z)\, \pi_{i,s}(\mathrm{d}z) \ .$$

They are approximated by a Monte Carlo sum:

$$\mathsf{h}_i(\widehat{S}_{t,k-\ell}) \approx \widehat{\mathsf{h}}_i^{t,k-\ell} \stackrel{\text{def}}{=} \frac{1}{m_{t,k+1}} \sum_{r=1}^{m_{t,k+1}} \mathcal{H}_i(Z_r^{i,t,k-\ell}), \ell \in \{0,1\}$$

where conditionally to the past of the algorithm $\mathcal{F}_{t,k}$, the random variables $\{Z_r^{i,t,k-\ell}, r \geq 0\}$ are independent and identically distributed (i.i.d.) with distribution $\pi_{i,\widehat{S}_{t,k-\ell}}$.

Section 4 provides an example of this setting. More precisely, $\mathcal{F}_{t,k}$ is the filtration associated to the history of the algorithm up to the outer loop #t and the inner loop #k,

$$\mathcal{F}_{t,-1} \stackrel{\text{def}}{=} \mathcal{F}_{t-1,k_{\mathrm{in}}} \ , \quad \mathcal{F}_{t,0} \stackrel{\text{def}}{=} \mathcal{F}_{t,-1} \bigvee \sigma(\widehat{\mathsf{h}}_i^{t,-1}, i \in [n]^\star) \ ,$$
$$\mathcal{F}_{t,k} \stackrel{\text{def}}{=} \mathcal{F}_{t,k-1} \bigvee \sigma\left(\mathcal{B}_{t,k}; \widehat{\mathsf{h}}_i^{t,k-1}, \widehat{\mathsf{h}}_i^{t,k-2}, i \in \mathcal{B}_{t,k}\right) \ .$$

For any $t \in [k_{\mathrm{out}}]^\star$ and $k \in [k_{\mathrm{in}} - 1]$, define

$$\eta_{t,k+1} \stackrel{\text{def}}{=} \frac{1}{\mathsf{b}} \sum_{i \in \mathcal{B}_{t,k+1}} \left(\widehat{\mathsf{h}}_i^{t,k} - \widehat{\mathsf{h}}_i^{t,k-1} - \mathsf{h}_i(\widehat{S}_{t,k}) + \mathsf{h}_i(\widehat{S}_{t,k-1})\right) \ ;$$

$\eta_{t,k+1}$ corresponds to the errors when approximating the quantities $\mathsf{h}_i(\widehat{S}_{t,k-\ell})$ for $\ell \in \{0,1\}$ at outer loop #t and inner loop #$(k+1)$. From standard computations on i.i.d. random variables (the randomness being the selection of the mini-batch $\mathcal{B}_{t,k+1}$), we have

$$\mathbb{E}[\eta_{t,k+1}|\mathcal{F}_{t,k}] = 0 \ , \tag{5}$$

$$\mathbb{E}\left[\|\eta_{t,k+1} - \mathbb{E}[\eta_{t,k+1}|\mathcal{F}_{t,k}]\|^2|\mathcal{F}_{t,k}\right] \leq \frac{C_v}{\mathsf{b}\,m_{t,k+1}} \ , \tag{6}$$

where

$$C_v \stackrel{\text{def}}{=} 2 \sup_{s \in \mathcal{S}} n^{-1} \sum_{i=1}^n \int_{\mathsf{Z}} \|\mathcal{H}_i(z) - \mathsf{h}_i(s)\|^2 \pi_{i,s}(\mathrm{d}z) \ .$$

The result (5) claims that the errors $\eta_{t,k+1}$ are unbiased, while (6) is the control of its conditional variance - which is a decreasing function of the batch size and the number of points in the Monte Carlo sum. In (6), the control is uniform with respect to the past (the right

hand side is deterministic while the left hand side is random): this assumption can be difficult to check when the optimization problem is not constrained in order to ensure that the points $\widehat{S}_{t,k}$ remain in a bounded subset of $\mathbb{R}^q$.

Theorem 1 provides an upper bound on a control in expectation of the convergence of the algorithm. First, it controls the difference of two successive values $\widehat{S}_{t,k} - \widehat{S}_{t,k-1}$; then it controls the quantity

$$\Delta_{t,k} \stackrel{\text{def}}{=} \frac{\left\|\mathrm{Prox}_{B(\widehat{S}_{t,k}),\gamma_{t,k}g}\left(\widehat{S}_{t,k-1} + \gamma_{t,k}\mathsf{h}(\widehat{S}_{t,k-1})\right) - \widehat{S}_{t,k-1}\right\|^2}{\gamma_{t,k}^2}$$

which is a kind of distance to the set of the solutions to (3) (see (4)). When $B(s) = \mathrm{I}_d$ and $g = 0$, $\Delta_{t,k} = \|\mathsf{h}(\widehat{S}_{t,k})\|^2 = \|\nabla \mathrm{W}(\widehat{S}_{t,k})\|^2$. The proof of Theorem 1 is given in [18].

Theorem 1 *Assume A1, A2 and A3. For any $t \in [k_{\mathrm{out}}]^\star$ and $k \in [k_{\mathrm{in}} - 1]$, set*

$$\gamma_{t,k} = \gamma_\star \stackrel{\text{def}}{=} \frac{v_{\min}}{L_{\dot{\mathrm{W}}} + 2Lv_{\max}\sqrt{k_{\mathrm{in}}}/\sqrt{\mathsf{b}}} \ .$$

Denote by (τ, K) a uniform random variable on $[k_{\mathrm{out}}]^\star \times [k_{\mathrm{in}}]$, independent of the path $\{\widehat{S}_{t,k}, t \in [k_{\mathrm{out}}]^\star, k \in [k_{\mathrm{in}}]\}$. Then,

$$\frac{v_{\min}^2}{2\left(L_{\dot{\mathrm{W}}} + 2Lv_{\max}\sqrt{k_{\mathrm{in}}}/\sqrt{\mathsf{b}}\right)} \mathbb{E}\left[\frac{\|\widehat{S}_{\tau,K} - \widehat{S}_{\tau,K-1}\|^2}{\gamma_{\tau,K}^2}\right]$$
$$\leq \frac{1}{k_{\mathrm{out}}(1 + k_{\mathrm{in}})} \left(\mathrm{W}(\widehat{S}_{\mathrm{init}}) + g(\widehat{S}_{\mathrm{init}}) - \min(\mathrm{W} + g)\right)$$
$$+ C_v \frac{v_{\max}}{2L} \frac{1}{\sqrt{k_{\mathrm{in}}}\mathsf{b}} \mathbb{E}\left[\frac{k_{\mathrm{in}} - K}{m_{\tau,K+1}}\right] \ .$$

In addition

$$\left(\frac{2}{v_{\min}}\left\{L_{\dot{\mathrm{W}}} + 2Lv_{\max}\sqrt{\frac{k_{\mathrm{in}}}{\mathsf{b}}}\right\} + L\sqrt{\frac{k_{\mathrm{in}}}{\mathsf{b}}}\right)^{-1} \mathbb{E}[\Delta_{\tau,K}]$$
$$\leq \left\{\frac{L_{\dot{\mathrm{W}}}}{Lv_{\min}} + 2\frac{v_{\max}}{v_{\min}}\sqrt{\frac{k_{\mathrm{in}}}{\mathsf{b}}}\right\}^{-1} \left(\frac{1}{L} + \left(\frac{v_{\max}}{v_{\min}}\right)^2 \gamma_\star\sqrt{\frac{k_{\mathrm{in}}}{\mathsf{b}}}\right) \cdots$$
$$\times \mathbb{E}\left[\frac{\|\widehat{S}_{\tau,K} - \widehat{S}_{\tau,K-1}\|^2}{\gamma_{\tau,K}^2}\right]$$
$$+ \left(\frac{v_{\max}}{v_{\min}}\right)^2 \frac{C_v}{L} \frac{1}{\sqrt{\mathsf{b}k_{\mathrm{in}}}} \mathbb{E}\left[\frac{k_{\mathrm{in}} - K}{m_{\tau,K}}\right] \ .$$

In Theorem 1, the expectations are w.r.t. the stochastic path of the algorithm $\{\widehat{S}_{t,k}\}$ and to the randomness of the times (τ, K). This theorem provides a control of the errors when the algorithm is stopped at some random time (τ, K). Such a control is classical in the non-convex case to show non-asymptotic convergence of stochastic gradient methods to a stationary point [19]: it consists in fixing a maximal number of iterations (here set to $k_{\mathrm{out}} \times k_{\mathrm{in}}$) of the algorithm, and draw at random, prior the run of the algorithm, a stopping time (τ, K).

Let us discuss the complexity bounds in the case $m_{t,k} = m$ for any $t \in [k_{\mathrm{out}}]^\star$ and $k \in [k_{\mathrm{in}}]^\star$. The total number of proximal calls is equal to $\mathcal{N}_P \stackrel{\text{def}}{=} k_{\mathrm{out}}(k_{\mathrm{in}} + 1)$. The total number of approximations of the functions h_i is equal to $\mathcal{N}_A \stackrel{\text{def}}{=} k_{\mathrm{out}}(n + 2\mathsf{b}k_{\mathrm{in}})$. The total number of Monte Carlo draws is $\mathcal{N}_{MC} \stackrel{\text{def}}{=} m\mathcal{N}_A$. Let us fix $\epsilon > 0$. Among the values $\mathcal{K}(n, \epsilon)$ of the positive integers

$(k_{\mathrm{out}}, k_{\mathrm{in}}, \mathsf{b}, m) \in \mathbb{N}^4$ which guarantee that ϵ-stationarity is reached i.e.

$$\mathbb{E}\left[\Delta_{\tau, K}\right] \le \epsilon \,,$$

the complexity bounds in terms of proximal calls are defined as $\mathcal{K}_P \stackrel{\mathrm{def}}{=} \min_{\mathcal{K}(n, \epsilon)} \mathcal{N}_P$; similarly, we define the complexity in terms of $\mathsf{h}_i's$ approximations, and in terms of Monte Carlo draws. By choosing $\mathsf{b} = k_{\mathrm{in}} = \sqrt{n}$, $k_{\mathrm{out}} = 1/(\sqrt{n}\epsilon)$ and $m = 1/\epsilon$, it is easily seen from Theorem 1 that $\mathbb{E}\left[\Delta_{\tau, K}\right] = O(\epsilon)$ and

$$\mathcal{K}_P = O(\epsilon^{-1})\,, \quad \mathcal{K}_A = O(\sqrt{n}\epsilon^{-1})\,, \quad \mathcal{K}_{MC} = O(\sqrt{n}\epsilon^{-2})\,.$$

When $B(s) = \mathrm{I}_q$, $g = 0$ and the gradient functions h_i's can be computed exactly, the state of the art complexity of variance-reduced gradient algorithm in terms of total number of computations of these gradient functions is $O(\sqrt{n}\epsilon^{-1})$ [8]. This bound is also reached by the variance reduced incremental EM named SPIDER-EM, which corresponds to the case $g = 0$, $B(s) \ne \mathrm{I}_d$, and the preconditioned gradient functions h_i's are explicit (see [15]). Our complexity $\mathcal{K}_A$ reaches this optimal value: in that sense, 3P-SPIDER is optimal since the bound $O(\sqrt{n}\epsilon^{-1})$ is reached despite the introduction of a proximal operator and the approximations of the preconditioned gradient functions h_i'. To reach this optimal bound, the Monte Carlo complexity is $O(\sqrt{n}\epsilon^{-2})$.

4. APPLICATION: INFERENCE IN THE LOGISTIC REGRESSION MODEL

We illustrate the convergence of 3P-SPIDER applied to inference in the following logistic regression model. Given n covariate vectors $\{X_i, i \in [n]^\star\}$ in $\mathbb{R}^d$ and $\theta \in \mathbb{R}^d$, the $\{-1, 1\}$-valued binary observations $\{Y_i, i \in [n]^\star\}$ are assumed independent with success probability $\mathbb{P}_\theta(Y_i = 1)$ equal to

$$\frac{1}{\sigma^d \sqrt{2\pi}^d} \int_{\mathbb{R}^d} (1 + \exp(-\langle X_i, z_i\rangle))^{-1} \exp\left(-\frac{\|z_i - \theta\|^2}{2\sigma^2}\right) \mathrm{d}z_i\,.$$

This model corresponds to a predictor Z_i for each individual #i and these predictors $Z_j, j \in [n]^\star$, are independent with distribution $\mathcal{N}_d(\theta, \sigma^2 \mathrm{I}_d)$. It is assumed that σ^2 is known; the unknown parameter θ is learnt by minimization of the penalized normalized negative log-likelihood $\theta \mapsto F(\theta)$, with penalty term $\mathrm{pen}(\theta) \stackrel{\mathrm{def}}{=} \tau\|\theta\|^2$ for some $\tau > 0$. F is equal to (see [1])

$$\theta \mapsto -\frac{1}{n}\sum_{i=1}^{n} \log \int_{\mathbb{R}} (1 + \exp(-Y_i\|X_i\|z))^{-1} \exp\left(\langle \mathsf{s}_i(z), \theta\rangle\right)$$
$$\times \exp(-z^2/(2\sigma^2))\mathrm{d}z + \mathsf{R}(\theta)$$

where $\mathsf{s}_i(z) \stackrel{\mathrm{def}}{=} zX_i/(\|X_i\|\sigma^2)$, $\mathsf{R}(\theta) \stackrel{\mathrm{def}}{=} (1/2)\theta^T\Omega^{-1}\theta$ and $\Omega \stackrel{\mathrm{def}}{=} \left(\frac{1}{\sigma^2 n}\sum_{i=1}^{n}\frac{X_iX_i^T}{\|X_i\|^2} + 2\tau\mathrm{I}_d\right)^{-1}$. The minimization of this criterion by a EM algorithm can be solved equivalently in the expectation space in order to minimize $\mathsf{W} : s \mapsto F(\Omega s)$ (see e.g. [20, 1]). In that case, EM finds the roots on $\mathbb{R}^d$ of

$$s \mapsto \mathsf{h}(s) \stackrel{\mathrm{def}}{=} \frac{1}{n}\sum_{i=1}^{n}\int_{\mathbb{R}} \mathsf{s}_i(z)p_i(z; \Omega s)\mathrm{d}z - s\,;$$

$z \mapsto p_i(z; \theta)$ is the probability density function proportional to

$$z \mapsto (1 + \exp(-Y_i\|X_i\|z))^{-1} \exp\left(\langle \mathsf{s}_i(z), \theta\rangle - z^2/(2\sigma^2)\right)\,.$$

We have $\nabla\mathsf{W}(s) \stackrel{\mathrm{def}}{=} -\Omega\mathsf{h}(s)$ for any $s \in \mathbb{R}^d$ (see [1]). Furthermore, upon noting that $\mathbb{P}_\theta(Y_i = y_i) \le 1$, it can be shown that the minima of F are in the set $\{\theta \in \mathbb{R}^d : \tau\|\theta\|^2 \le \ln 4\}$, which implies that EM in the expectation space will find the roots of h in $\mathcal{K} \stackrel{\mathrm{def}}{=} \{s \in \mathbb{R}^d : s^T\Omega s \le \ln 4/(\tau\lambda_{\min})\}$ where $\lambda_{\min}$ is the positive minimal eigenvalue of Ω. Therefore, we set g equal to the characteristic function of $\mathcal{K}$; with such a definition of $\mathcal{K}$, the associated weighted proximal is explicit.

The data set is built from the MNIST data set, as described in [1, Section 3]: $n = 24\,989$ and $d = 51$. SPIDER-EM is run with $\sigma^2 = 0.1$, $\tau = 1$, $k_{\mathrm{out}} = 20$, $k_{\mathrm{in}} = \lceil\sqrt{n}/10\rceil = 16$, $\mathsf{b} = \lceil 10\sqrt{n}\rceil$, $\mathcal{E}_t = 0$, $\gamma_{t,0} = 0$, $\gamma_{t,k} = 0.1$ and $m_{t,k} = 2\lceil\sqrt{n}\rceil$ until the outer loop #9 and then $m_{t,k} = 10\lceil\sqrt{n}\rceil$.

On Figure 1(a), the 51 components of the sequence $\{\widehat{S}_{t,k_{\mathrm{in}}}, t \in [k_{\mathrm{out}}]^\star\}$ are displayed vs the index of the outer loop t. The convergence can be observed.

On Figure 1(b), we display the quantiles 0.25, 0.5 and 0.75 of the squared norm $\|\widehat{S}_{t,k}\|^2$ as a function of the cumulated number of inner loops; these quantiles are computed over 25 independent runs of 3P-SPIDER. Here again, the convergence and the stability of the path over the independent runs can be observed.

Finally, Figures 1(c,d) display the quantiles 0.25 and 0.75 of $\widehat{\Delta}_{t,k} \stackrel{\mathrm{def}}{=} \|\widehat{S}_{t,k} - \widehat{S}_{t,k-1}\|^2/\gamma_{t,k}^2$ as a function of the cumulated number of inner loops; these quantiles are estimated over 25 independent runs of 3P-SPIDER. We observe the gain when increasing the number of Monte Carlo points in order to reduce the fluctuations of the approximations of the h_i's; see [1, Section 3] for a detailed study of the design parameters of 3P-SPIDER. To illustrate the benefit of the variance-reduction step in 3P-SPIDER, we also run Prox-Online-EM with $\mathsf{b} = \lceil 10\sqrt{n}\rceil$ and $m_{t,k} = 2\lceil\sqrt{n}\rceil$. Prox-Online-EM corresponds to Online-EM combined with a proximal step i.e. a proximal-preconditioned gradient algorithm. The quantiles 0.25 and 0.75 of $\|\widehat{S}_t - \widehat{S}_{t-1}\|^2/\gamma_t^2$ are displayed on Figures 1(c,d) as a function of the number of iterations t. It illustrates that 3P-SPIDER, as a proximal variance-reduced preconditioned gradient method, clearly improves on Online-EM.
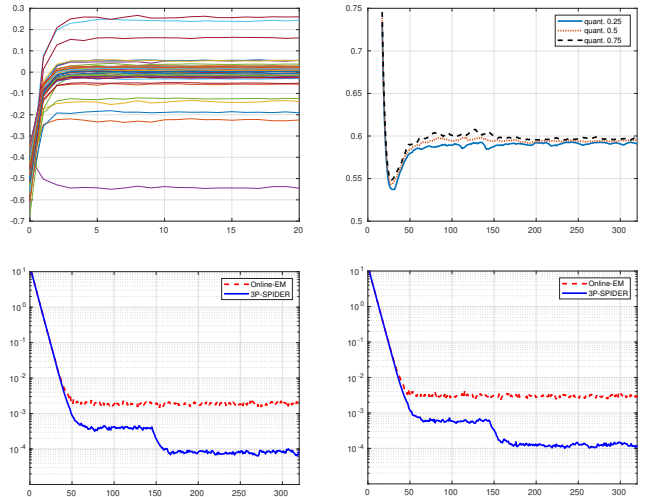

Fig. 1. [(a) top left] Sequence $\{\widehat{S}_{t,k_{\mathrm{in}}}, t \in [k_{\mathrm{out}}]^\star\}$ [(b) top right] Quantiles of $\|\widehat{S}_{t,k}\|^2$ [(c) bottom left] Quantile 0.25 of $\widehat{\Delta}_{t,k}$ [(d) bottom right] Quantile 0.75 of $\widehat{\Delta}_{t,k}$

5. REFERENCES

[1] G. Fort and E. Moulines, "The Perturbed Prox-Preconditioned SPIDER algorithm for EM-based large scale learning," in *2021 IEEE Statistical Signal Processing Workshop (SSP)*, 2021, Accepted.

[2] L. Bottou, F. E Curtis, and J. Nocedal, "Optimization methods for large-scale machine learning," *Siam Review*, vol. 60, no. 2, pp. 223–311, 2018.

[3] R. Johnson and T. Zhang, "Accelerating stochastic gradient descent using predictive variance reduction," *Advances in neural information processing systems*, vol. 26, pp. 315–323, 2013.

[4] M. Schmidt, N. Le Roux, and F. Bach, "Minimizing finite sums with the stochastic average gradient," *Mathematical Programming*, vol. 162, no. 1-2, pp. 83–112, 2017.

[5] Z. Allen-Zhu, "Natasha 2: Faster non-convex optimization than sgd," in *Advances in Neural Information Processing Systems*, S. Bengio, H. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, and R. Garnett, Eds. 2018, vol. 31, Curran Associates, Inc.

[6] Z. Allen-Zhu and Y. Li, "Neon2: Finding local minima via first-order oracles," in *NeurIPS*, 2018, pp. 3720–3730.

[7] S.J. Reddi, S. Sra, B. Poczos, and A.J. Smola, "Proximal stochastic methods for nonsmooth nonconvex finite-sum optimization," in *Proceedings of the 30th International Conference on Neural Information Processing Systems*, 2016, pp. 1153–1161.

[8] Z. Wang, K. Ji, Y. Zhou, Y. Liang, and V. Tarokh, "SpiderBoost and Momentum: Faster Stochastic Variance Reduction Algorithms," in *Advances in Neural Information Processing Systems 32*, H. Wallach, H. Larochelle, A. Beygelzimer, F. d' Alché-Buc, E. Fox, and R. Garnett, Eds., pp. 2406–2416. 2019.

[9] R. Johnson and T. Zhang, "Accelerating Stochastic Gradient Descent using Predictive Variance Reduction," in *Advances in Neural Information Processing Systems 26*, C. J. C. Burges, L. Bottou, M. Welling, Z. Ghahramani, and K. Q. Weinberger, Eds., pp. 315–323. Curran Associates, Inc., 2013.

[10] A. Beck and M. Teboulle, "Gradient-based algorithms with applications to signal-recovery problems," in *Convex Optimization in Signal Processing and Communications*, 2010.

[11] 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.

[12] N. Parikh and S. Boyd, "Proximal algorithms," *Found. Trends Optim.*, vol. 1, no. 3, pp. 127–239, 2014.

[13] L. M. Nguyen, J. Liu, K. Scheinberg, and M. Takáč, "Sarah: A novel method for machine learning problems using stochastic recursive gradient," in *Proceedings of the 34th International Conference on Machine Learning - Volume 70*. 2017, ICML'17, p. 2613–2621, JMLR.org.

[14] C. Fang, C.J. Li, Z. Lin, and T. Zhang, "SPIDER: Near-Optimal Non-Convex Optimization via Stochastic Path-Integrated Differential Estimator," in *Advances in Neural Information Processing Systems 31*, S. Bengio, H. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, and R. Garnett, Eds., pp. 689–699. Curran Associates, Inc., 2018.

[15] G. Fort, E. Moulines, and H.-T. Wai, "A stochastic path integral differential estimator expectation maximization algorithm," in *Advances in Neural Information Processing Systems*, H. Larochelle, M. Ranzato, R. Hadsell, M. F. Balcan, and H. Lin, Eds. 2020, vol. 33, pp. 16972–16982, Curran Associates, Inc.

[16] Y.F. Atchadé, G. Fort, and E. Moulines, "On perturbed proximal gradient algorithms," *Journal of Machine Learning Research*, vol. 18, no. 10, pp. 1–33, 2017.

[17] G. Fort, P. Gach, and E. Moulines, "Fast Incremental Expectation Maximization for finite-sum optimization: nonasymptotic convergence," *Statistics and Computing*, 2021, Accepted for publication.

[18] G. Fort and E. Moulines, "Perturbed Prox-Preconditioned SPIDER for finite-sum optimization," Tech. Rep., 2021, work in progress.

[19] S. Ghadimi and G. Lan, "Stochastic First- and Zeroth-Order Methods for Nonconvex Stochastic Programming," *SIAM J. Optimiz.*, vol. 23, no. 4, pp. 2341–2368, 2013.

[20] B. Delyon, M. Lavielle, and E. Moulines, "Convergence of a Stochastic Approximation version of the EM algorithm," *Ann. Statist.*, vol. 27, no. 1, pp. 94–128, 1999.