# Stochastic Variable Metric Proximal Gradient with variance reduction for non-convex composite optimization

Gersende Fort, Eric Moulines

# Stochastic Variable Metric Proximal Gradient with variance reduction for non-convex composite optimization

Gersende Fort[1*] and Eric Moulines[2]

[1*]Institut de Mathématiques de Toulouse, CNRS & Université de Toulouse, 118 route de Narbonne, Toulouse, 31400, France.
[2]CMAP, Ecole Polytechnique, Route de Saclay, Palaiseau, 91128 Cedex, France.

*Corresponding author(s). E-mail(s): gersende.fort@math.univ-toulouse.fr;
Contributing authors: eric.moulines@polytechnique.edu;

**Abstract**

This paper introduces a novel algorithm, the Perturbed Proximal Preconditioned SPIDER algorithm (`3P-SPIDER`), designed to solve finite sum non-convex composite optimization. It is a stochastic Variable Metric Forward-Backward algorithm, which allows approximate preconditioned forward operator and uses a variable metric proximity operator as the backward operator; it also proposes a mini-batch strategy with variance reduction to address the finite sum setting. We show that `3P-SPIDER` extends some Stochastic preconditioned Gradient Descent-based algorithms and some Incremental Expectation Maximization algorithms to composite optimization and to the case the forward operator can not be computed in closed form. We also provide an explicit control of convergence in expectation of `3P-SPIDER`, and study its complexity in order to satisfy the epsilon-approximate stationary condition. Our results are the first to combine the non-convex composite optimization setting, a variance reduction technique to tackle the finite sum setting by using a minibatch strategy and, to allow deterministic or random approximations of the preconditioned forward operator. Finally, through an application to inference in a logistic regression model with random effects, we numerically compare `3P-SPIDER` to other stochastic forward-backward algorithms and discuss the role of some design parameters of `3P-SPIDER`.

**Keywords:** Stochastic optimization, Variable Metric Forward-Backward splitting, Preconditioned Stochastic Gradient Descent, Incremental Expectation Maximization, Proximal methods, Variance reduction, Non-asymptotic convergence bounds.

## 1 Introduction

Efficient learning from large data sets require new optimization algorithms designed to be robust to big data and complex models era. In Statistics and Machine Learning, we are often faced with solving problems of the form

$$\mathrm{argmin}_{s\in\mathbb{R}^q}\left(\frac{1}{n}\sum_{i=1}^{n} W_i(s) + g(s)\right) ,$$

where $n$ is the number of examples in the training data set, $s$ is an unknown quantity to be learnt from the examples, $W_i$ is a loss function associated to the example $\#i$ and $g$ is a regularization

term encoding a priori knowledge and constraints on $s$; $g$ may also prevent from overfitting. Quite often, the regularization term $g : \mathbb{R}^q \to (0, +\infty]$ is not differentiable, and the data fidelity term $n^{-1} \sum_{i=1}^n W_i$ is smooth on the domain of $g$.

This paper is concerned with stochastic optimization of a non-convex finite sum composite function; more precisely, it addresses the differential inclusion problem

$$0 \in \frac{1}{n} \sum_{i=1}^{n} G_i(s) + \partial g(s), \qquad s \in \mathbb{R}^q, \qquad (1)$$

where $g : \mathbb{R}^q \to (-\infty, +\infty]$ is lower semi-continuous convex with non-empty domain $\mathcal{S}$ and for all $i$, $G_i : \mathcal{S} \to \mathbb{R}^q$ is globally Lipschitz on $\mathcal{S}$.

The first goal of this paper is to provide a novel algorithm. Motivated by applications in Statistics and Machine learning, we require this algorithm to satisfy the following three constraints. (c1) The algorithm uses possibly preconditioned operators $\mathsf{h}_i$ instead of the forward operator $G_i$:

$$\forall s \in \mathcal{S}, \qquad \mathsf{h}_i(s, B) \stackrel{\text{def}}{=} -B^{-1} G_i(s), \qquad (2)$$

where $B$ is a $q \times q$ positive definite matrix. Such a condition encompasses preconditioned gradient methods for example, which also includes gradient methods with adaptive step sizes. It also encompasses Expectation Maximization (EM) algorithms (Dempster et al (1977)) designed for large scale learning. (c2) The algorithm may only have access to approximations of $\mathsf{h}_i(s, B)$. Such a condition addresses the situations when $\mathsf{h}_i(s, B)$ is not explicit, for example when it is defined by an intractable integral. This occurs at each E-step of EM, when the conditional expectations under the a posteriori distributions can not be computed exactly. (c3) The algorithm addresses the finite sum challenge while keeping the caused variability induced by the algorithmic solution as small as possible. For example, when the solution relies on a random selection of a mini-batch of examples, the algorithm has to propose a variance reduction scheme.

A first class of problems of the form (1) are minimizations of regularized loss functions through gradient-based algorithms. In that case, $G_i \stackrel{\text{def}}{=} \nabla W_i$. (2) allows preconditioned gradients; such a variable metric is known to accelerate the convergence. Variable Metric Forward-Backward (VMFB) algorithms were introduced to solve (1)-(2) in the case $G$ is a gradient. Nevertheless, as discussed in Section 2.1, to our best knowledge none of the variants of VMFB address the three constraints c1, c2 and c3 simultaneously.

A second application of (1)-(2) is the EM algorithm, an algorithm originally designed to compute the Maximum-Likelihood estimator of an unknown parameter $\theta$ in latent variable models. When the complete data model is from the curved exponential family, EM is equivalent to an algorithm in the so-called *statistic space* (see e.g. Delyon et al (1999)). This remark is the cornerstone of stochastic EM algorithms including incremental EM ones designed for incremental processing of large data sets. EM only supplies preconditioned forward operators $-\mathsf{B}(s)^{-1} G_i(s)$. Therefore, stochastic EM algorithms are naturally in the setting (1)-(2) (see Fort et al (2020)). Here again, as discussed in Section 2.2, none of the EM variants in the literature address the constraints (c2) and (c3) simultaneously.

Our first contribution is the design of a novel iterative algorithm, named `Perturbed Proximal Preconditioned SPIDER` (`3P-SPIDER`), which combines *(i)* a preconditioned forward operator associated to the smooth part $n^{-1} \sum_{i=1}^n G_i(s)$, *(ii)* a variable metric proximity operator with respect to the non-smooth part $g$, *(iii)* a stochastic approach to address the finite sum setting induced by $n^{-1} \sum_{i=1}^n G_i(s)$ combined with a variance reduction technique based on the `SPIDER` algorithm (Nguyen et al (2017); Fang et al (2018)); it also allows *(iv)* numerical approximations of the preconditioned forward operator when it has no closed form. The algorithm is introduced in Section 3, together with discussions on implementation questions. We also design a stochastic VMFB algorithm which answers the constraints c1 and c2 but does not contain a variance reduction step as required by c3.

The second contribution is to provide a non-asymptotic convergence analysis in expectation of `3P-SPIDER` in the case the variable metric $B$ at iteration $\#t$ depends on the current value of the iterate, and the $G_i$'s are gradient operators; see Section 4. The proof relies on a Lyapunov inequality with an original construction, which

is a consequence of the non-convex optimization setting, and the fact the algorithm uses preconditioned forward operators and variable metric proximity operators (see Section 7.4).
Theorem 4.1 provides a control of convergence in expectation for 3P-SPIDER, which explicitly identifies the impact of non-exact preconditioned forward operators, and the impact of initialization strategies. First, we prove that the learning rate of 3P-SPIDER can be chosen constant over iterations when the preconditioned forward operator is exact or replaced with an unbiased random oracle; and is decreasing along iterations when it is replaced with a biased oracle (deterministic or random). Second, we provide the first convergence result for a stochastic VMFB algorithm addressing c1, c2 and c3 for non-convex finite sum composite optimization. For example, it is the first result for incremental EM with a non-smooth penalty term ($g \neq 0$) and possibly biased Monte Carlo approximations of the E-step.
When the forward operator $\mathsf{h}_i(s, \mathsf{B}(s))$ is exact, we study the complexity of 3P-SPIDER (see Corollary 4.3): in order to satisfy the $\epsilon$-approximate stationary condition, the number of calls $\mathcal{K}_{\bar{\mathsf{h}}}$ to one of the operator $\mathsf{h}_i(s, \mathsf{B}(s))$ is $O(\sqrt{n}/\epsilon)$, the number of calls $\mathcal{K}_{\text{prox}}$ to the backward operator is $O(1/\epsilon)$ and, the learning rate can be chosen independent of the accuracy $\epsilon$. Applied to the Gradient method and applied to the EM method when there are no constraints ($g = 0$), these explicit controls of convergence retrieve previous results in the literature (see e.g. Wang et al (2019); Fort et al (2020)) which are known to be at the state-of-the-art.
Finally we show that this complexity analysis remains valid when the forward operators are approximated. In the difficult case when the approximations are biased random oracles based on Monte Carlo sums, we show that $\mathcal{K}_{\bar{\mathsf{h}}}$ and $\mathcal{K}_{\text{prox}}$ are not impacted by the approximation and are the same as with exact operators $\mathsf{h}_i(\cdot, \mathsf{B}(\cdot))$, by choosing an adequate number of terms in the Monte Carlo sums. The price to be paid is a Monte Carlo complexity of order $O(\sqrt{n}/\epsilon^2)$.

In Section 5, 3P-SPIDER is applied to inference in a logistic regression model with random effects. We show how the problem is of the form (1)-(2). In this example, the preconditioned forward operators are approximated by a Monte Carlo sum computed from a Markov chain Monte Carlo sampler. Through numerical analyses in the case 3P-SPIDER is a stochastic Expectation Maximization algorithm in the statistic space, we discuss the choice of design parameters. We also show how the SPIDER variance reduction effect can be increased by exploiting the Monte Carlo approximations of the forward operator.

The proofs are given in Section 6 and Section 7; technical details are also provided in Appendix.

**Notations.** We denote by $\langle \cdot, \cdot \rangle$ the dot product on $\mathbb{R}^q$, and by $\| \cdot \|$ the associated norm. For a $q \times q$ positive definite matrix $B$, we set $\langle \cdot, \cdot \rangle_B \stackrel{\text{def}}{=} \langle B \cdot, \cdot \rangle$ and $\| \cdot \|_B$ the associated norm. $\mathrm{I}_q$ denotes the $q \times q$ identity matrix. For a matrix $B$, $B^\top$ is its transpose. $\mathcal{P}_+^q$ denotes the set of the $q \times q$ positive definite matrices.

$\mathbb{N}$ (resp. $\mathbb{N}^\star$) is the set of non negative (resp. positive) integers. For $n \in \mathbb{N}^\star$, we set $[n] \stackrel{\text{def}}{=} \{0, \cdots, n\}$ and $[n]^\star \stackrel{\text{def}}{=} \{1, \cdots, n\}$. $\mathbb{R}_+$ is the set of the positive real numbers. For $q \in \mathbb{R}$, $\lceil q \rceil$ is the upper integer part.

I is the identity function. For a proper function $g : \mathbb{R}^q \to (-\infty, +\infty]$, $\partial g(s)$ denotes the subdifferential at $s$. For a continuously differentiable function $W$ at $s \in \mathbb{R}^q$, $\nabla W(s)$ is the gradient of $W$ at $s$.

All the random variables (r.v.) are defined on $(\Omega, \mathcal{A}, \mathbb{P})$; for a r.v. $U$, $\sigma(U)$ is the sigma-field generated by $U$.

# 2 Motivating examples

In this section, we show that the Variable Metric Proximal-Gradient algorithm and the Expectation Maximization algorithm are examples of the general framework described by (1) and (2). In the first case, the preconditioning matrices $B$ are chosen by the user, while in the second case, they are supplied by the algorithm.

## 2.1 Variable Metric Proximal-Gradient algorithms

Consider the non-convex composite problem with finite sum structure

$$\text{find } s \in \mathbb{R}^q: \quad 0 \in \frac{1}{n} \sum_{i=1}^{n} \nabla W_i(s) + \partial g(s) \,, \qquad (3)$$

where $g$ is a proper lower semicontinuous convex function with domain $\mathcal{S}$; and for all $i \in [n]^\star$, $W_i$ is continuously differentiable on $\mathcal{S}$. It is of the form (1) with $G_i \overset{\text{def}}{=} \nabla W_i$ being a gradient. (3) is an example of the more general problem: finding a zero on $\mathbb{R}^q$ of the sum of two (set-valued) operators $0 \in \mathsf{A}s + \mathsf{C}s$. Here, $\mathsf{A} \overset{\text{def}}{=} n^{-1} \sum_{i=1}^n \nabla W_i$ and $\mathsf{C} \overset{\text{def}}{=} \partial g$ is a maximally monotone operator (see e.g. (Bauschke and Combettes, 2011, Theorem 20.25)). (3) can be solved by Forward-Backward splitting algorithms (see e.g. Combettes and Wajs (2005); Beck (2017)): the forward step uses the gradient of some if not all the functions $W_i$'s at each iteration; the backward step uses a proximity operator associated to $g$. This yields Proximal-Gradient based algorithms.

In the case $g = 0$, which includes unconstrained optimization problems, stochastic gradient methods with variance reduction were proposed in the situation

$$n^{-1} \sum_{i=1}^{n} G_i(s) = \mathbb{E}\left[\mathcal{G}(Z, s)\right] \tag{4}$$

and random oracles $\mathcal{G}(Z, s)$ are available; in the non-convex setting, let us cite e.g. Ghadimi and Lan (2013); Reddi et al (2016); Allen-Zhu and Hazan (2016); Nguyen et al (2017); Allen-Zhu (2018); Fang et al (2018); Zhou et al (2020). These algorithms address the problem (1)-(2) by choosing $B$ equal to the identity matrix $\mathrm{I}_q$ and they use unbiased random oracles $\mathcal{G}(Z, s)$ for the approximation of the forward operator.

For non-convex composite optimization ($g \neq 0$), let us cite Ghadimi et al (2016) and Karimi et al (2016) for stochastic Proximal-Gradient algorithms using unbiased oracles $\mathcal{G}(Z, s)$ (see (4)). Li and Li (2018), Wang et al (2019), Zhang and Xiao (2019), Nhan et al (2020) and Metel and Takeda (2021) propose stochastic Proximal-Gradient methods with unbiased random oracles and including variance reduction schemes. In Metel and Takeda (2021), $g$ may be non-convex but admits an efficiently computable proximity operator. Atchadé et al (2017) allow for deterministic or random approximations of the forward operator $n^{-1} \sum_{i=1}^n G_i(s)$; when the perturbation is stochastic, the convergence analysis covers both biased and unbiased oracles, includes Nesterov acceleration schemes, but is restricted to convex optimization. Here again, all these algorithms address the problem (1)-(2) by choosing $B = \mathrm{I}_q$.

Forward-Backward suffers from slow convergence, and Variable Metric Forward-Backward (`VMFB`) methods were proposed by Chen and Rockafellar (1997) in order to accelerate the convergence (see also refs. 11 to 16 in Chouzenoux et al (2014)). `VMFB` changes the metric at each iteration by using symmetric positive definite scaling matrices multiplying the forward operator. It is an alternative to inertial methods such as Heavy Ball or Nesterov acceleration which use informations from the previous iterates. When solving the inclusion (3), `VMFB` uses preconditioned gradients with an iteration-dependent preconditioning matrix $B_t^{-1}$ for the forward step at iteration $\#t$, and scales the proximal step consequently. Examples showing that `VMFB` is more efficient than Forward-Backward and Forward-Backward with inertial schemes, are provided in Chouzenoux et al (2014) and Repetti et al (2014). Different strategies exist for the definition of the variable metric $B_t$; for example, it may be a diagonal matrix depending on the past history of the algorithm (see e.g. Park et al (2019) and references therein for variable scalar metrics; see also Chen et al (2019) in the case $g = 0$), or inherited from Newton-type methods (see e.g. Becker and Fadili (2012); Lee et al (2014); Becker et al (2019) for the composite convex case; see also Kolte et al (2015); Moritz et al (2016); Gower et al (2016) for the smooth convex case with finite sum structure; finally, see Zhang et al (2022) for the smooth non-convex case with finite sum structure), or defined through a Majorize-Minimize strategy to make the backward operator explicit (see e.g. Chouzenoux et al (2014) and Repetti and Wiaux (2021)). Convergence results for `VMFB` exist in the convex case (see e.g. Combettes and Vũ (2014) and Bonettini et al (2021); and Park et al (2019) for the strongly convex case) and in the non-convex case (see e.g. Chouzenoux et al (2014) and Repetti and Wiaux (2021)). In Yun et al (2021), a stochastic `VMFB` is studied in the non-convex case; the exact gradient is approximated by a linear combination of random oracles, with exponential forgetting, and the oracles are assumed unbiased and bounded.

`3P-SPIDER` addresses non-convex composite optimization with a finite sum structure by using Proximal-Gradient algorithms accelerated via the

Variable Metric cunning. It is a stochastic VMFB, which contains a variance reduction technique in order to overcome the finite sum setting; it also allows oracles for the preconditioned forward operators, oracles which can be biased or unbiased when random (see e.g. Atchadé et al (2017) and Fort et al (2018) for examples motivating biased random approximations of the gradient). The combination of these two sources of perturbations is an original setting which, to our best knowledge, is not addressed in the literature.

3P-SPIDER uses preconditioning matrices, which may depend on the current value of the iterate and therefore may be random. The non-asymptotic convergence analysis derived in Section 4 will rely on weaker minorization assumptions on the spectrum of the scaling matrices (see A 3) than in Yun et al (2021); it will not require ordering assumptions on the sequence of scaling matrices as in Combettes and Vũ (2014) and Bonettini et al (2021), and will not assume a Kurdyka-Łojasiewicz condition on the objective function as in Chouzenoux et al (2014). As a consequence, the construction of the Lyapunov inequality for the convergence analysis of 3P-SPIDER differs from these previous works.

3P-SPIDER requires the backward operator to be explicit, which may be a strong assumption especially for variable metric proximity operators (see Section 3.5); extensions of the convergence analysis to the case of inexact proximity operators is out of the scope of this paper.

## 2.2 Expectation Maximization for curved exponential families

Consider the parametric statistical model: the observations are independent with density

$$y \mapsto \int_{\mathsf{Z}} p(y, z;\ \theta) \mu_{lv}(\mathrm{d}z) \ ,$$

with respect to (w.r.t.) a $\sigma$-finite positive measure $\mu_o$ on $\mathbb{R}^{d_y}$. In this model, $z$ acts as a *latent* variable taking values in the measurable set $(\mathsf{Z}, \mathcal{Z})$ endowed with a $\sigma$-finite positive measure $\mu_{lv}$ (see e.g. Everitt (1984) for examples of latent variable models). The goal is to learn the parameter $\theta \in \Theta \subseteq \mathbb{R}^d$ from $n$ observations $Y_1, \cdots, Y_n$, by minimizing the negative normalized log-likelihood

$$F(\theta) \stackrel{\text{def}}{=} -\frac{1}{n} \sum_{i=1}^{n} \log \int_{\mathsf{Z}} p(Y_i, z;\ \theta) \mu_{lv}(\mathrm{d}z) \qquad (5)$$

on $\Theta$. Unfortunately, this is a non-convex problem and most often, an optimization algorithm for the minimization of (5) can not do better than converging to a critical point of the objective function (see Wu (1983)).

A popular model is the case when the *complete data likelihood* $(y, z) \mapsto p(y, z;\ \theta)$ is of the form

$$p(y, z;\ \theta) \stackrel{\text{def}}{=} H(y, z) \exp\left(\langle S(y, z), \phi(\theta) \rangle - \psi(\theta)\right)$$

where $H : \mathbb{R}^{d_y} \times \mathsf{Z} \to \mathbb{R}_+$, $S : \mathbb{R}^{d_y} \times \mathsf{Z} \to \mathbb{R}^q$, $\phi : \Theta \to \mathbb{R}^q$, $\psi : \Theta \to \mathbb{R}$; it corresponds to the so-called *curved exponential family* assumption. It is satisfied by the mixture models as soon as the components of the mixture are from the curved exponential family. See e.g. Brown (1986) for an introduction to curved exponential family of distributions; and McLachlan and Krishnan (2008) for examples of such latent variable models.

EM for curved exponential families defines iteratively a $\Theta$-valued sequence $\{\theta_t, t \geq 0\}$ through the mechanism: given $\theta_t$,

- **(E-step)** Compute $\bar{s}(\theta_t)$, the expectation of the *sufficient statistics* w.r.t. the *a posteriori* distributions

$$\bar{s}(\theta_t) \stackrel{\text{def}}{=} \frac{1}{n} \sum_{i=1}^{n} \bar{s}_i(\theta_t) \ ,$$
$$\bar{s}_i(\theta_t) \stackrel{\text{def}}{=} \int_{\mathsf{Z}} S(Y_i, z) \frac{p(Y_i, z;\ \theta_t)\, \mu_{lv}(\mathrm{d}z)}{\int_{\mathsf{Z}} p(Y_i, u;\ \theta_t) \mu_{lv}(\mathrm{d}u)} \ .$$

- **(M-step)** Update the parameter

$$\theta_{t+1} \stackrel{\text{def}}{=} \operatorname{argmin}_{\theta \in \Theta} \left(\psi(\theta) - \langle \bar{s}(\theta_t), \phi(\theta) \rangle\right) \ .$$

The algorithm alternates between a step in the parameter space $\Theta$ (when computing $\theta_{t+1} \in \mathbb{R}^d$), and a step in the *statistic space* when computing $\bar{s}(\theta_t) \in \mathbb{R}^q$. Proposition 2.1 states that the limiting points of EM run in the parameter space $\Theta$ are the fixed points of an operator onto $\Theta$; finding such a fixed point is equivalent to find a fixed point of an operator onto the statistic space $\bar{s}(\Theta) \subseteq \mathbb{R}^q$.

**Proposition 2.1** *Assume that for any* $s \in \mathcal{S} \supseteq \bar{s}(\Theta)$,

$$\mathsf{T}(s) \stackrel{\text{def}}{=} \operatorname{argmin}_{\theta\in\Theta} \left(\psi(\theta) - \langle s, \phi(\theta)\rangle\right)$$

*exists and is unique. Set* $\mathcal{L}_\theta \stackrel{\text{def}}{=} \{\theta \in \Theta : \mathsf{T}(\bar{s}(\theta)) = \theta\}$ *and* $\mathcal{L}_s \stackrel{\text{def}}{=} \{s \in \bar{s}(\Theta) : \bar{s}(\mathsf{T}(s)) = s\}$.

*$\mathcal{L}_\theta$ is the set of the limiting points of EM. If $\theta_\star \in \mathcal{L}_\theta$, then $s_\star \stackrel{\text{def}}{=} \bar{s}(\theta_\star)$ is in $\mathcal{L}_s$. Conversely, if $s_\star \in \mathcal{L}_s$ then $\theta_\star \stackrel{\text{def}}{=} \mathsf{T}(s_\star)$ is in $\mathcal{L}_\theta$.*

See e.g. Delyon et al (1999) for the proof. An algorithmic corollary of Proposition 2.1, is that EM is equivalent to any algorithm run in the statistic space and designed to find the roots of the function

$$s \mapsto \frac{1}{n}\sum_{i=1}^{n} \bar{\mathsf{h}}_i(s) \;, \quad \text{where} \quad \bar{\mathsf{h}}_i(s) \stackrel{\text{def}}{=} \bar{s}_i(\mathsf{T}(s)) - s \;,$$

on the subset $\bar{s}(\Theta)$ of $\mathbb{R}^q$. Under regularity conditions on the statistical model, it is proved in Delyon et al (1999) (see also a statement in (Fort et al, 2020, Proposition 1)) that there exists a $q\times q$ positive definite matrix $\mathsf{B}(s)$ such that

$$\nabla\left(F \circ \mathsf{T}\right)(s) = -\mathsf{B}(s)\left(\frac{1}{n}\sum_{i=1}^{n} \bar{\mathsf{h}}_i(s)\right) \;, \qquad (6)$$

where $F$ is the negative normalized log-likelihood (see (5)). Therefore, the roots of $n^{-1}\sum_{i=1}^n \bar{\mathsf{h}}_i(s)$ on $\bar{s}(\Theta)$ are the roots of $n^{-1}\sum_{i=1}^n G_i(s)$ on $\bar{s}(\Theta)$, where $G_i(s) \stackrel{\text{def}}{=} -\mathsf{B}(s)\,\bar{\mathsf{h}}_i(s)$. It also means that the roots of $n^{-1}\sum_{i=1}^n \bar{\mathsf{h}}_i(s)$ are the roots of the gradient of $F\circ\mathsf{T}$, the objective function transferred on the statistic space through the map $\mathsf{T} : \Theta \to \mathbb{R}^q$.

As a conclusion, EM in the statistic space is an example of problem (1)-(2), where the function $g$ collects the constraint on $s$ such as $s \in \mathcal{S} \supseteq \bar{s}(\Theta)$: *(i)* it is designed to find a root of $n^{-1}\sum_{i=1}^n G_i(s) = \nabla(F\circ\mathsf{T})(s)$ under the constraint that $s \in \mathcal{S}$; *(ii)* it uses the quantities $\bar{\mathsf{h}}_i(s)$ which are preconditioned forward operator since there exists $B(s)$ such that $\bar{\mathsf{h}}_i(s) = -\mathsf{B}(s)^{-1}\,G_i(s)$ (see (6)); *(iii)* this preconditioned forward operator is intractable for at least two reasons: first, due to the inner integrations on the set $\mathsf{Z}$ when computing $\bar{s}_i$, since $p(y,z;\,\theta)$ and $\mathsf{Z}$ are often too complex to make the integrals explicit; second, due to the outer integration on the $n$ examples when computing $\bar{s}$, which has a prohibitive computational cost in large scale learning. However, a Monte Carlo approximation of $\bar{\mathsf{h}}_i(s)$ is always possible, whatever $i$ and $s$. This remark is the cornerstone to understand the stochastic versions of EM (see e.g. Celeux and Diebolt (1985); Wei and Tanner (1990); Delyon et al (1999); Fort and Moulines (2003) which address the inner sum intractability; and Neal and Hinton (1998); Ng and McLachlan (2003); Cappé and Moulines (2009); Chen et al (2018); Karimi et al (2019); Fort et al (2020, 2021a) for the outer sum intractability). They consist in running a Stochastic Approximation (SA) algorithm with mean field $n^{-1}\sum_{i=1}^n \bar{\mathsf{h}}_i(s)$ (for an introduction to SA, see e.g. Benveniste et al (1990) or Borkar (2008)); this yields *SA within EM* procedures. They differ through the construction of the random field used for the approximation of the mean field (see (Fort et al, 2021a, Section 2.2.) for a description of some SA within EM algorithms).

`3P-SPIDER` is among the SA within EM algorithms. Compared to previous stochastic EM methods, it encompasses the two random approximations (of the sum in $i$ and of the integrals on $\mathsf{Z}$) and a variance reduction step, and it also allows a more general penalty term $g$ than the $\{0,+\infty\}$-valued indicator function of a set.

# 3 The `3P-SPIDER` algorithm

We introduce a novel algorithm named *Perturbed Proximal Preconditioned SPIDER* (`3P-SPIDER`), solving (1) and satisfying (c1), (c2) and (c3). It requires $g$ to satisfy the following assumption

**A1.** *$g : \mathbb{R}^q \to (-\infty,+\infty]$ is proper, lower semi-continuous and convex. Denote by $\mathcal{S}$ its domain $\mathcal{S} \stackrel{\text{def}}{=} \{s \in \mathbb{R}^q : g(s) < +\infty\}$.*

Under this assumption, we define a *variable metric* proximity operator. For any $\gamma > 0$ and $B \in \mathcal{P}_+^q$, the proximity operator of the proper lower semicontinuous convex function $\gamma g : \mathbb{R}^q \to (-\infty,+\infty]$ relative to the metric induced by $B$ is defined by (see e.g. (Hiriart-Urruty and Lemaréchal, 1996, Section XV.4))

$$\operatorname{prox}_{\gamma g}^{B}(s) \stackrel{\text{def}}{=} \operatorname{argmin}_{\mathbb{R}^q}\left(\gamma g(\cdot) + \frac{1}{2}\|\cdot - s\|_B^2\right). \quad (7)$$

When $B = \mathrm{I}_q$, we simply write $\mathrm{prox}_{\gamma g}(s)$, which is the proximity operator originally defined by Moreau (1965). Lemma 3.1 shows that under A1, $\mathrm{prox}^B_{\gamma g}(s)$ exists and is unique for all $s \in \mathbb{R}^q$, $\gamma > 0$ and $B \in \mathcal{P}^q_+$. It also provides characterizations of this point. Its proof is in Section 6.1.

**Lemma 3.1** *Assume A1.*

1. *For any $\gamma > 0$, $B \in \mathcal{P}^q_+$ and $s \in \mathbb{R}^q$, the optimization problem (7) has a unique solution, characterized as the unique point $\mathsf{p} \in \mathcal{S}$ satisfying*

$$-\gamma^{-1} B\,(\mathsf{p} - s) \in \partial g(\mathsf{p})\ .$$

2. *For any $\gamma > 0$, $B \in \mathcal{P}^q_+$, $s \in \mathcal{S}$ and $h \in \mathbb{R}^q$,*

$$s = \mathrm{prox}^B_{\gamma g}(s + \gamma h) \quad \textit{iff} \quad Bh \in \partial g(s). \tag{8}$$

For $s \in \mathbb{R}^q$ and $B \in \mathcal{P}^q_+$, set

$$\mathsf{h}_i(s, B) \stackrel{\text{def}}{=} -B^{-1} G_i(s)\ ,$$
$$\mathsf{h}(s, B) \stackrel{\text{def}}{=} n^{-1} \sum_{i=1}^n \mathsf{h}_i(s, B)\ . \tag{9}$$

By Lemma 3.1-item 2, it holds for any $B \in \mathcal{P}^q_+$ and $\gamma > 0$: $s = \mathrm{prox}^B_{\gamma g}\,(s + \gamma\,\mathsf{h}(s, B))$ iff $-B\mathsf{h}(s, B) \in \partial g(s)$. By (9), this yields for any $B \in \mathcal{P}^q_+$, and $\gamma > 0$:

$$s_\star = \mathrm{prox}^B_{\gamma g}\,(s_\star + \gamma\ \mathsf{h}(s_\star, B)) \quad \text{iff } s_\star \text{ solves (1)}. \tag{10}$$

## 3.1 Variable Metric Proximal and Preconditioned Gradient

(10) shows that when solving the composite optimization problem (1), as soon as a preconditioned version of the operator $s \mapsto n^{-1} \sum_{i=1}^n G_i(s)$ is used – with preconditioning matrix $B^{-1}$, a proximity operator of $g$ relative to a metric induced by the matrix $B$ has to be used.

Based on the characterization (10), a natural splitting algorithm to solve (1) under the condition (c1) is: given $\sigma_0 \in \mathcal{S}$, a positive stepsize sequence $\{\gamma_{k+1}, k \geq 0\}$ and a $\mathcal{P}^q_+$-valued sequence $\{B_{k+1}, k \geq 0\}$, repeat

$$\sigma_{k+1} = \mathrm{prox}^{B_{k+1}}_{\gamma_{k+1} g}\,(\sigma_k + \gamma_{k+1} \mathsf{h}(\sigma_k, B_{k+1}))\ . \tag{11}$$

It corresponds to the Variable Metric Forward-Backward algorithm (see e.g. Chen and Rockafellar (1997); Combettes and Vũ (2014)).

In the large scale learning setting, the full sum over the $n$ functions $\mathsf{h}_i$ (see (9)) can not be computed at each iteration of (11). In addition, it may happen that $\mathsf{h}_i(s)$ is not explicit (see e.g. the case of the incremental EM algorithms, Section 2.2). Therefore, a natural idea is to propose the inexact version of (11) defined by Algorithm 1: the proximal step is unchanged (see line 8); the SA step in line 7 uses a random approximation $\mathsf{S}_{k+1}$ of the exact mean field $n^{-1} \sum_{i=1}^n \mathsf{h}_i(\widehat{S}_k)$; this approximation, defined by line 6, combines a mini-batch approximation of a full sum (see line 3) and possibly approximated terms $\delta_{k+1,i}$ (see line 5).

**Algorithm 1** A stochastic Variable Metric Forward-Backward

**Require:** $k^{\mathrm{out}} \in \mathbb{N}^\star$, $\gamma_k > 0$ for $k \in [k^{\mathrm{out}}]^\star$, $\mathsf{b} \in \mathbb{N}^\star$, $\widehat{S}_{\mathrm{init}} \in \mathcal{S}$.
**Ensure:** The sequence $\{\widehat{S}_k, k \in [k^{\mathrm{out}}]\}$.

1: $\widehat{S}_0 = \widehat{S}_{\mathrm{init}}$
2: **for** $k = 0, \cdots, k^{\mathrm{out}} - 1$ **do**
3: Sample a batch $\mathcal{B}_{k+1}$ of size $\mathsf{b}$ in $[n]^\star$
4: Choose $B_{k+1} \in \mathcal{P}^q_+$
5: For $i \in \mathcal{B}_{k+1}$, compute an approximation $\delta_{k+1,i}$ of $\mathsf{h}_i(\widehat{S}_k, B_{k+1})$.
6: $\mathsf{S}_{k+1} = \mathsf{b}^{-1} \sum_{i \in \mathcal{B}_{k+1}} \delta_{k+1,i}$
7: $\widehat{S}_{k+1/2} = \widehat{S}_k + \gamma_{k+1}\ \mathsf{S}_{k+1}$
8: $\widehat{S}_{k+1} = \mathrm{prox}^{B_{k+1}}_{\gamma_{k+1} g}(\widehat{S}_{k+1/2})$.
9: **end for**

## 3.2 The SPIDER variance reduction technique

3P-SPIDER leverages on Algorithm 1 and on the variance reduction technique SPIDER for the definition of the field $\mathsf{S}_{k+1}$ that approximates $n^{-1} \sum_{i=1}^n \mathsf{h}_i(\widehat{S}_k, B_{k+1})$. SPIDER stands for *Stochastic Path-Integrated Differential EstimatoR*, and was originally introduced in the stochastic gradient descent literature by Fang et al (2018) (see also Nguyen et al (2017); Wang et al (2019)).

We give the intuition of SPIDER in the SA setting which encompasses the stochastic gradient one.

SA scheme solves a root finding problem $\xi(s) = 0$ on $\mathbb{R}^q$ by: given an initial value $s_0 \in \mathbb{R}^q$ and a stepsize sequence $\{\gamma_{k+1}, k \geq 0\}$, repeat $s_{k+1} = s_k + \gamma_{k+1}\, \Xi_{k+1}$, where at each iteration $\#(k+1)$, $\Xi_{k+1}$ is a random approximation of $\xi(s_k)$. Usually, it is required that conditionally to the past of the algorithm, the expectation of $\Xi_{k+1}$ is $\xi(s_k)$; in that case, $\Xi_{k+1}$ can be replaced with $\mathsf{S}_{k+1} \stackrel{\text{def}}{=} \Xi_{k+1} + V_{k+1}$, where conditionally to the past, $V_{k+1}$ is centered. SPIDER leverages on this remark and on the *control variate* technique: it proposes a clever construction of a random variable $V_{k+1}$ approximating zero and correlated to $\Xi_{k+1}$.

The recipe is as follows: consider that at iteration $\#k$, $\mathsf{S}_k$ is a random approximation of $\mathsf{h}(\widehat{S}_{k-1}, B_k)$. Then define $\mathsf{S}_{k+1}$ by $\mathsf{S}_{k+1} \stackrel{\text{def}}{=} H_{k+1} + V_{k+1}$ where

$$H_{k+1} \stackrel{\text{def}}{=} \mathsf{b}^{-1} \sum_{i \in \mathcal{B}_{k+1}} \mathsf{h}_i(\widehat{S}_k, B_{k+1}) \;,$$

$$V_{k+1} \stackrel{\text{def}}{=} \mathsf{S}_k - \mathsf{b}^{-1} \sum_{i \in \mathcal{B}_{k+1}} \mathsf{h}_i(\widehat{S}_{k-1}, B_k) \;,$$

and $\mathcal{B}_{k+1}$ is sampled at random in $[n]^\star$. The r.v. $V_{k+1}$ approximates zero since both $\mathsf{b}^{-1} \sum_{i \in \mathcal{B}_{k+1}} \mathsf{h}_i(\widehat{S}_{k-1}, B_k)$ and $\mathsf{S}_k$ approximate $n^{-1} \sum_{i=1}^n \mathsf{h}_i(\widehat{S}_{k-1}, B_k)$; $V_{k+1}$ and $H_{k+1}$ are correlated via $\mathcal{B}_{k+1}$.

Unfortunately, the r.v. $\mathsf{S}_{k+1}$ is not an unbiased approximation of $n^{-1} \sum_{i=1}^n \mathsf{h}_i(\widehat{S}_k, B_{k+1})$ (see Proposition 7.3 in the case $B_{k+1}$ is of the form $\mathsf{B}(\widehat{S}_k)$). In order to remove the bias, SPIDER restarts the control variate mechanism regularly: every $k^{\text{in}}$ iterations, compute a full sum over the $n$ terms and set $\mathsf{S}_{k^{\text{in}}+1} = n^{-1} \sum_{i=1}^n \mathsf{h}_i(\widehat{S}_{k^{\text{in}}}, B_{k^{\text{in}}+1})$.

### 3.3 3P-SPIDER

3P-SPIDER is given by Algorithm 2. The iteration index is $(t, k)$ where $t$ is the index of the current *outer* loop and ranges from 1 to $k^{\text{out}}$, and $k$ is the index of the current *inner* loop. At outer loop $\#t$, there are $k_t^{\text{in}}$ inner iterations. The inner iterations are Algorithm 1 (see Lines 8, 9, 12 and 13 of Algorithm 2) combined with the SPIDER variance reduction trick (see Line 11 of Algorithm 2) adapted to the case when the quantities $\mathsf{h}_i(\widehat{S}_{t,k}, B_{t,k+1}) - \mathsf{h}_i(\widehat{S}_{t,k-1}, B_{t,k})$ can not be computed exactly (see Line 10).

When $G_i$ is a gradient and $B(s) = \mathrm{I}_q$, different strategies were proposed for SPIDER for the choice of $\mathsf{b}'_t$ and $k_t^{\text{in}}$. In Fang et al (2018); Nguyen et al (2017); Wang et al (2019), the number of inner loops is constant ($k_t^{\text{in}} = k^{\text{in}}$ for any $t \geq 1$) and $\mathsf{b}'_t = n$; Nguyen et al (2017) also considers the case when $k_t^{\text{in}}$ is adapted based on the history of the algorithm while being upper bounded; in Horváth et al (2022), $\mathsf{b}'_t$ is deterministic and depends on $t$, $\mathsf{b}$ depends on $t$, and $k_t^{\text{in}}$ is a Geometric random variable with an expectation depending on $t$; in Li et al (2021), $\mathsf{b}'_t$ does not depend on $t$ and $k_t^{\text{in}}$ is random.

For the EM problem (see Section 2.2), Fort et al (2020) introduced SPIDER-EM, a variance reduced stochastic EM designed for large scale learning, in a situation when the computation of $\bar{\mathsf{h}}_i(s)$ is exact for all $s, i$. For this algorithm, the benefit of an increasing batch size $t \mapsto \mathsf{b}'_t$ and a geometric number of inner loops $k_t^{\text{in}}$ with time-varying expectation, is discussed in Fort et al (2021b). The conclusion is that the best strategy is a deterministic increasing sequence $\mathsf{b}'_t$ in order to have an increasing accuracy when refreshing the variable $\mathsf{S}_\cdot$, and a constant number of inner loops $k_t^{\text{in}} = k^{\text{in}}$. This paper allows $\mathsf{b}'_t$ and $k_t^{\text{in}}$ to vary with $t$: they may be deterministic functions of $t$ or random ones as well.

The matrices $\{B_{t,k+1}, t \in [k^{\text{out}}]^\star, k \in [k^{\text{in}} - 1]\}$ can be deterministic or random. They could be chosen prior the run of the algorithm; more efficient strategies consist in adapting this matrix along the run of the algorithm, based on its history. In EM (see Section 2.2), $B_{t,k+1}$ is of the form $\mathsf{B}(\widehat{S}_{t,k})$ where $\mathsf{B}$ is defined by the statistic model.

After $k_t^{\text{in}}$ inner iterations, the outer loop $\#(t+1)$ starts: the stochastic mean field $\mathsf{S}_{t+1,0}$ is refreshed (see Line 3 to Line 6). Here again, two approximations of the original SPIDER algorithm are allowed: the first one is when computing $\mathsf{h}_i(\widehat{S}_{t,0}, B_{t,1})$ and the second one avoids the scan of the full data set (one may choose $\mathsf{b}'_t < n$).

The input variables of 3P-SPIDER are the number of outer loops $k^{\text{out}}$, the number of inner loops $k_t^{\text{in}}$, the stepsize sequence $\{\gamma_{t,k}, t \geq 1, k \geq 1\}$ for the SA steps, the size of the mini-batches $\mathsf{b}$ and

**Algorithm 2** The Perturbed Proximal Preconditioned SPIDER algorithm (`3P-SPIDER`)

**Require:** $k^{\text{out}} \in \mathbb{N}^\star$, $k_t^{\text{in}} \in \mathbb{N}^\star$ for $t \in [k^{\text{out}}]^\star$, $\gamma_{t,k+1} > 0$ for $t \in [k^{\text{out}}]^\star, k \in [k_t^{\text{in}}]$, $\mathsf{b} \in \mathbb{N}^\star$, $\mathsf{b}'_t \in \mathbb{N}^\star$ for $t \in [k^{\text{out}}]^\star$, $\widehat{S}_{\text{init}} \in \mathcal{S}$ and $B_{\text{init}} \in \mathcal{P}_+^q$

**Ensure:** The sequence $\{\widehat{S}_{t,k}, t \in [k^{\text{out}}]^\star, k \in [k_t^{\text{in}}]^\star\}$.

$\widehat{S}_{0,k_0^{\text{in}}} = \widehat{S}_{\text{init}}$, $B_{0,k_0^{\text{in}}} = B_{\text{init}}$
**for** $t = 1, \cdots, k^{\text{out}}$ **do**
  $\widehat{S}_{t,0} = \widehat{S}_{t-1,k_{t-1}^{\text{in}}}$, $\widehat{S}_{t,-1} = \widehat{S}_{t-1,k_{t-1}^{\text{in}}}$, $B_{t,0} = B_{t-1,k_{t-1}^{\text{in}}}$
  Sample a batch $\mathcal{B}_{t,0}$ of size $\mathsf{b}'_t$ in $[n]^\star$, with or without replacement.
  For all $i \in \mathcal{B}_{t,0}$, compute $\delta_{t,0,i}$ equal to or approximating $\mathsf{h}_i(\widehat{S}_{t,0}, B_{t,0})$.
  $\mathsf{S}_{t,0} = (\mathsf{b}'_t)^{-1} \sum_{i \in \mathcal{B}_{t,0}} \delta_{t,0,i}$
  **for** $k = 0, \cdots, k_t^{\text{in}} - 1$ **do**
    Sample a mini batch $\mathcal{B}_{t,k+1}$ of size $\mathsf{b}$ in $[n]^\star$, with or without replacement.
    Choose $B_{t,k+1} \in \mathcal{P}_+^q$.
    For all $i \in \mathcal{B}_{t,k+1}$, compute $\delta_{t,k+1,i}$ equal to or approximating $\mathsf{h}_i(\widehat{S}_{t,k}, B_{t,k+1}) - \mathsf{h}_i(\widehat{S}_{t,k-1}, B_{t,k})$.
    $\mathsf{S}_{t,k+1} = \mathsf{S}_{t,k} + \mathsf{b}^{-1} \sum_{i \in \mathcal{B}_{t,k+1}} \delta_{t,k+1,i}$
    $\widehat{S}_{t,k+1/2} = \widehat{S}_{t,k} + \gamma_{t,k+1}\, \mathsf{S}_{t,k+1}$
    $\widehat{S}_{t,k+1} = \text{prox}_{t,k}(\widehat{S}_{t,k+1/2})$, where $\text{prox}_{t,k} \overset{\text{def}}{=} \text{prox}_{\gamma_{t,k+1} g}^{B_{t,k+1}}$.
  **end for**
**end for**

$\mathsf{b}'_t$, and the initial values of the iterate $\widehat{S}_{\text{init}}$ and the metric $B_{\text{init}}$ in $\mathcal{S}$ and $\mathcal{P}_+^q$ respectively.

## 3.4 Monte Carlo approximation of $\mathsf{h}_i(s, B)$

Set $\vartheta \overset{\text{def}}{=} (s, i, B) \in \mathcal{S} \times [n]^\star \times \mathcal{P}_+^q$. In some applications, there exist a measurable function $H_\vartheta$ and a probability measure $\pi_\vartheta$ defined on the measurable set $(\mathsf{Z}, \mathcal{Z})$ such that

$$\mathsf{h}_i(s, B) = \int_{\mathsf{Z}} H_\vartheta(z) \pi_\vartheta(\mathrm{d}z) \,. \tag{12}$$

This is the case of EM in the statistic space (see Section 2.2) where $H_\vartheta(z) = S(Y_i, z) - s$ and

$$\pi_\vartheta(\mathrm{d}z) \overset{\text{def}}{=} \frac{p(Y_i, z;\ \mathsf{T}(s))}{\int_{\mathsf{Z}} p(Y_i, u;\ \mathsf{T}(s))\, \mu_{lv}(\mathrm{d}u)}\, \mu_{lv}(\mathrm{d}z) \,.$$

When the integral in (12) is intractable, one can resort to Monte Carlo integrations to define the approximations $\delta_{t,k+1,i}$ and $\delta_{t,0,i}$ (see e.g. Devroye (1986) for exact sampling methods, and Robert and Casella (2004) for an introduction to Markov chain Monte Carlo methods). If $\{Z_m^\vartheta, m \geq 0\}$ are independent samples with distribution $\pi_\vartheta(\mathrm{d}z)$ or are a path of an ergodic Markov chain with unique invariant distribution $\pi_\vartheta(\mathrm{d}z)$, then we can set

$$\mathsf{h}_i(\widehat{S}_{t,k}, B_{t,k+1}) \approx \frac{1}{M} \sum_{m=1}^{M} H_{\vartheta_{t,k+1,i}}(Z_m^{\vartheta_{t,k+1,i}}) \,,$$

where $\vartheta_{t,k+1,i} \overset{\text{def}}{=} (\widehat{S}_{t,k}, i, B_{t,k+1})$. We will show numerically in Section 5 that when approximating the difference $\mathsf{h}_i(\widehat{S}_{t,k}, B_{t,k+1}) - \mathsf{h}_i(\widehat{S}_{t,k-1}, B_{t,k})$, there is a gain in correlating the two sequences $\{Z_m^{\vartheta_{t,k+1,i}}, m \geq 0\}$ and $\{Z_m^{\vartheta_{t,k,i}} m \geq 0\}$; this makes stronger the effect of the `SPIDER` control variate (see Section 3.2).

## 3.5 The computation of $\text{prox}_{\gamma g}^B$

When $g = 0$, $\text{prox}_{\gamma g}^B(s) = s$. When $g \neq 0$, $\mathsf{p} \overset{\text{def}}{=} \text{prox}_{\gamma g}^B(s)$ solves $0 \in \mathsf{p} - s + \gamma B^{-1} \partial g(\mathsf{p})$ and there does not always exist an explicit expression of $\mathsf{p}$.

When $B = \mathrm{I}_q$, (Combettes and Pesquet, 2011, Tables 10.1 and 10.2) provide properties of $\text{prox}_{\gamma g}$ and expressions of proximity operators for many functions $g$.

When $B$ is the sum of a diagonal matrix and of a rank one matrix, (Becker and Fadili, 2012, Section 3) presents iterative algorithms for the computation of $\mathsf{p}$. For a general positive definite matrix $B$, we have from (Combettes and Vũ, 2014,

Example 3.9)

$$\mathrm{prox}^{B}_{\gamma g}(s) = \sqrt{B}^{-1}\mathrm{prox}_{\gamma g(\sqrt{B}^{-1}\cdot)}(\sqrt{B}s)\ ,$$

where $\sqrt{B}$ is the square root of the matrix $B$. (Becker and Fadili, 2012, Lemma 5) (see also Combettes and Vũ (2014)) establishes a Moreau identity i.e. an expression of $\mathrm{prox}^{B}_{\gamma g}$ as a function of a proximity operator of the Fenchel conjugate of $g$.

In the special case $g$ is the $\{0,+\infty\}$-valued indicator function of a closed convex set $\mathcal{S}$, the *projected Landweber* method is an iterative algorithm for the computation of $\mathsf{p}$ (see Eicke (1992), see also (Combettes and Pesquet, 2011, Example 10.10)).

Finally, for applicatons including a metric selection step, metric selection strategies for the definition of $B$ can be found in (Park et al, 2019, Section 3) for diagonal variable metrics; and in Repetti and Wiaux (2021) for specific functions $g$ which circumvent the often challenging computation of $\mathrm{prox}^{B}_{\gamma g}$.

# 4 Non-asymptotic convergence analysis

This section is devoted to explicit non-asymptotic bounds for the convergence in expectation of `3P-SPIDER`. We will restrict to the case there exist $\mathsf{B}:\mathcal{S}\to\mathcal{P}^q_+$ and

$$B_{t,k+1} \stackrel{\mathrm{def}}{=} \mathsf{B}(\widehat{S}_{t,k})\ .$$

This framework encompasses the EM problem (see Section 2.2) and any preconditioned gradient-based algorithms (see Section 2.1) when the preconditioning matrix depends on the past history of the algorithm via the current value of the iterate. We will also use the notation

$$\bar{\mathsf{h}}_i(s) \stackrel{\mathrm{def}}{=} \mathsf{h}_i(s,\mathsf{B}(s))\ ,\quad \bar{\mathsf{h}}(s) \stackrel{\mathrm{def}}{=} \mathsf{h}(s,\mathsf{B}(s))\ . \quad (13)$$

`3P-SPIDER` is designed to solve (1) under the constraints `c1` to `c3`. Therefore, based on (10), we are interested in a control of the quantities $\mathrm{prox}_{t,k}\left(\widehat{S}_{t,k}+\gamma_{t,k+1}\,\bar{\mathsf{h}}(\widehat{S}_{t,k})\right)-\widehat{S}_{t,k}$ where

$$\mathrm{prox}_{t,k}(s) \stackrel{\mathrm{def}}{=} \mathrm{prox}^{\mathsf{B}(\widehat{S}_{t,k})}_{\gamma_{t,k+1}g}(s)\ .$$

Roughly speaking, these quantities evaluate how far the algorithm is from the limiting set at iteration $\#(t,k)$. More precisely, we will control the cumulative "distances to stationary" $\sum_{t=1}^{k^{\mathrm{out}}}\sum_{k=0}^{k_t^{\mathrm{in}}-1}\Delta^{\star}_{t,k+1}$ where $\Delta^{\star}_{t,k+1}$ is equal to

$$\frac{\|\mathrm{prox}_{t,k}(\widehat{S}_{t,k}+\gamma_{t,k+1}\bar{\mathsf{h}}(\widehat{S}_{t,k}))-\widehat{S}_{t,k}\|^2_{\mathsf{B}(\widehat{S}_{t,k})}}{\gamma^2_{t,k+1}}\ ; \quad (14)$$

$\mathsf{h}$ is defined by (9).

The controls in expectation of the cumulated distances are obtained under the assumptions A 2 to A 4. A 2 is a smoothness assumption on the functions $\mathsf{h}_i$, A 3 assumes that $n^{-1}\sum_{i=1}^n G_i(s)$ is a gradient operator of some so-called *Lyapunov function*, and the spectrum of the matrices $\mathsf{B}(s)$ are bounded uniformly in $s$. A 4 are assumptions on the approximations $\delta_{t,k+1,i}$.

**A 2.** *For all $i\in[n]^{\star}$, the function $\bar{\mathsf{h}}_i$ is globally Lipschitz on $\mathcal{S}$, with constant $L_i$: there exists a positive constant $L_i$ such that $\forall s,s'\in\mathcal{S}$, $\|\bar{\mathsf{h}}_i(s)-\bar{\mathsf{h}}_i(s')\|\le L_i\,\|s-s'\|$. Set $L^2\stackrel{\mathrm{def}}{=}n^{-1}\sum_{i=1}^n L_i^2$.*

A 2 only requires a Lipschitz property on $\mathcal{S}$; it is weaker than assuming the Lipschitz property on the full space $\mathbb{R}^q$ as sometimes assumed in the literature (see e.g. Combettes and Wajs (2005)). A 2 holds for example when $\mathcal{S}$ is compact and for all $i\in[n]^{\star}$, the gradient $\nabla\mathsf{h}_i$ exists and is continuous on $\mathcal{S}$.

**A 3.** *1. There exists a function $\mathrm{W}:\mathbb{R}^q\to\mathbb{R}$, continuously differentiable on $\mathcal{S}$ and such that*

$$\forall s\in\mathcal{S},\qquad \nabla\mathrm{W}(s)=\frac{1}{n}\sum_{i=1}^n G_i(s)\ ;$$

*in addition, $\bar{\mathsf{h}}_i(s) = -\mathsf{B}(s)^{-1}G_i(s)$, where $\mathsf{B}(s)\in\mathcal{P}^q_+$.*

*2. $\nabla\mathrm{W}$ is globally $L_{\dot{\mathrm{W}}}$-Lipschitz on $\mathcal{S}$.*

*3. There exist $0<v_{\min}\le v_{\max}<+\infty$ such that for any $s\in\mathcal{S}$, $v_{\min}\|\cdot\|^2\le\|\cdot\|^2_{\mathsf{B}(s)}\le v_{\max}\|\cdot\|^2$.*

Here again, both the Lipschitz property and the boundedness condition on the spectrum of the matrices $\mathsf{B}(s)$ are required on $\mathcal{S}$ and not on the full space $\mathbb{R}^q$. When $\mathsf{B}(s)$ does not depend on

$s$ ($\mathsf{B}(s) = B$ for any $s \in \mathcal{S}$), we have $L_{\dot{\mathrm{W}}} \leq v_{\max} n^{-1} \sum_{i=1}^{n} L_i$.

The last assumption is on the fluctuations of the errors when approximating $\bar{\mathsf{h}}_i(\widehat{S}_{t,k}) - \bar{\mathsf{h}}_i(\widehat{S}_{t,k-1})$: set $\xi_{t,k+1,i} \stackrel{\text{def}}{=} \delta_{t,k+1,i} - \bar{\mathsf{h}}_i(\widehat{S}_{t,k}) + \bar{\mathsf{h}}_i(\widehat{S}_{t,k-1})$ and define its conditional bias and variance, conditionally to the $\sigma$-field generated by $\mathcal{B}_{t,k+1}$, $\widehat{S}_{t,k}$ and $\widehat{S}_{t,k-1}$. Set $\mathcal{P}_{t,k+1/2} \stackrel{\text{def}}{=} \sigma\left(\mathcal{B}_{t,k+1}, \widehat{S}_{t,k}, \widehat{S}_{t,k-1}\right)$.

$$\mu_{t,k+1,i} \stackrel{\text{def}}{=} \mathbb{E}\left[\xi_{t,k+1,i} | \mathcal{P}_{t,k+1/2}\right]$$
$$\sigma^2_{t,k+1,i} \stackrel{\text{def}}{=} \mathbb{E}\left[\|\xi_{t,k+1,i} - \mu_{t,k+1,i}\|^2 | \mathcal{P}_{t,k+1/2}\right] .$$

We assume

**A4.** *1. Conditionally to $\mathcal{B}_{t,k+1}$, $\widehat{S}_{t,k}$ and $\widehat{S}_{t,k-1}$, the approximations $\{\delta_{t,k+1,i}, i \in \mathcal{B}_{t,k+1}\}$ are independent.*

*2. There exists a non negative constant $C_b$ and for any $t \in [k^{\text{out}}]^\star$, there exists a non decreasing deterministic sequence $\{m_{t,k}, k \geq 1\}$ such that for any $k \in [k_t^{\text{in}} - 1]$, with probability one,*

$$\|\frac{1}{n}\sum_{i=1}^{n} \mu_{t,k+1,i}\| \leq \frac{C_b}{m_{t,k+1}} .$$

*3. There exist non negative constants $C_v$ and $C_{vb}$ and for any $t \in [k^{\text{out}}]^\star$, there exist non decreasing deterministic sequences $\{M_{t,k}, k \geq 1\}$ and $\{\bar{M}_{t,k}, k \geq 1\}$ such that for any $k \in [k_t^{\text{in}} - 1]$, with probability one,*

$$\frac{1}{n}\sum_{i=1}^{n} \sigma^2_{t,k+1,i} \leq \frac{C_v}{M_{t,k+1}} ,$$
$$\frac{1}{n}\sum_{i=1}^{n} \|\mu_{t,k+1,i} - \frac{1}{n}\sum_{j=1}^{n} \mu_{t,k+1,j}\|^2 \leq \frac{C_{vb}^2}{\bar{M}^2_{t,k+1}} .$$

We allow the errors $\xi_{t,k+1,i}$ to be deterministic or random. When there are no errors ($\xi_{t,k+1,i} = 0$) then $C_b = C_v = C_{vb} = 0$. When the errors are deterministic, we have $\xi_{t,k+1,i} = \mu_{t,k+1,i}$ and $\sigma^2_{t,k+1,i} = 0$. When the errors are random and unbiased, then $\mu_{t,k+1,i} = 0$. Therefore, some of the constants $C_b$, $C_v$ or $C_{vb}$ can be null as summarized in Table 1.

In Section A, we discuss how A 4 is verified in the case $\bar{\mathsf{h}}_i(s') - \bar{\mathsf{h}}_i(s)$ is an expectation under a distribution that may depend on $(s, s', i)$ (see Section 3.4), and $\delta_{t,k+1,i}$ is a Monte Carlo approximation.

| | $C_b$ | $C_v$ | $C_{vb}$ |
|---|---|---|---|
| exact | 0 | 0 | 0 |
| deterministic | $\geq 0$ | 0 | $\geq 0$ |
| random, unbiased | 0 | $\geq 0$ | 0 |
| random, biased | $> 0$ | $\geq 0$ | $\geq 0$ |

**Table 1** The sign of the constants $C_b$, $C_v$, $C_{vb}$ when there are no approximations on the $\bar{\mathsf{h}}_i(s)'$ (case *exact*), and when there are approximations.

Theorem 4.1 provides an explicit upper bound of the cumulative distance to stationarity as measured by $\Delta^\star_{t,k+1}$ (see (14)) along the $\sum_{t=1}^{k^{\text{out}}} k_t^{\text{in}}$ iterations of the algorithm. It also provides an upper bound on the cumulative errors $\mathcal{D}^\star_{t,k+1}$ defined by

$$\frac{\|\widehat{S}_{t,k+1} - \text{prox}_{t,k}(\widehat{S}_{t,k} + \gamma_{t,k+1}\bar{\mathsf{h}}(\widehat{S}_{t,k}))\|^2_{\mathsf{B}(\widehat{S}_{t,k})}}{\gamma^2_{t,k+1}} ,$$

where $\bar{\mathsf{h}}$ is defined by (13). Given the current iterate $\widehat{S}_{t,k}$, $\mathcal{D}^\star_{t,k+1}$ compares two iterations: the ideal one $\text{prox}_{t,k}(\widehat{S}_{t,k} + \gamma_{t,k+1}\bar{\mathsf{h}}(\widehat{S}_{t,k}))$ and the available one $\text{prox}_{t,k}(\widehat{S}_{t,k} + \gamma_{t,k+1}\mathsf{S}_{t,k+1})$.

**Theorem 4.1** *Assume A 1, A 2, A 3 and A 4. Let $\{k_t^{\text{in}}, t \in [k^{\text{out}}]^\star\}$ be a deterministic positive sequence. For any $t \in [k^{\text{out}}]^\star$ and $k \in [k_t^{\text{in}} - 1]$, define $\Lambda_{t,k+1}$ by*

$$\frac{\gamma_{t,k} L_{\dot{\mathrm{W}}}}{v_{\min}} + \gamma^2_{t,k} L^2 \frac{2 v_{\max} k_t^{\text{in}}}{v_{\min} \mathsf{b}} \left(1 + \frac{2\, C_{vb}}{\sqrt{\mathsf{b}}\, \bar{M}_{t,k+1}}\right) . \quad (15)$$

*Let $\{\widehat{S}_{t,k}, t \in [k^{\text{out}}]^\star, k \in [k_t^{\text{in}}]^\star\}$ be the sequence given by Algorithm 2 when the stepsize sequence $\{\gamma_{t,k+1}, t \in [k^{\text{out}}]^\star, k \in [k^{\text{in}} - 1]\}$ satisfies*

$$\gamma_{t,k+1}\left(1 + \frac{2C_b}{m_{t,k+1}}\right) \leq \gamma_{t,k} , \quad \Lambda_{t,k+1} \in (0, 1/2) . \quad (16)$$

*Then,*

$$\sum_{t=1}^{k^{\text{out}}} \sum_{k=1}^{k_t^{\text{in}}} \gamma_{t,k} \left(\frac{1}{2} - \Lambda_{t,k+1}\right) \left\{\mathbb{E}\left[\Delta^\star_{t,k}\right] + \mathbb{E}\left[\mathcal{D}^\star_{t,k}\right]\right\}$$
$$\leq \mathbb{E}\left[\mathrm{W}(\widehat{S}_{1,0}) + g(\widehat{S}_{1,0})\right] - \min_{\mathcal{S}}(\mathrm{W} + g)$$
$$+ v_{\max} \sum_{t=1}^{k^{\text{out}}} \gamma_{t,0}\, k_t^{\text{in}}\, \mathbb{E}\left[\|\mathcal{E}_t\|^2\right]$$
$$+ v_{\max} \sum_{t=1}^{k^{\text{out}}} \sum_{k=1}^{k_t^{\text{in}}} \left(k_t^{\text{in}} - k + 1\right) \gamma_{t,k}\, \mathcal{U}_{t,k} ,$$

*where* $\mathcal{E}_t \overset{\text{def}}{=} \mathsf{S}_{t,0} - \mathsf{h}(\widehat{S}_{t,0})$ *and*

$$\mathcal{U}_{t,k} \overset{\text{def}}{=} \frac{2\,C_b}{m_{t,k}} + \frac{C_b^2}{m_{t,k}^2} + \frac{C_v}{\mathsf{b}\,M_{t,k}} + \frac{2\,C_{vb}}{\sqrt{\mathsf{b}}\,\bar{M}_{t,k}} + \frac{C_{vb}^2}{\mathsf{b}\,\bar{M}_{t,k}^2} \,.$$

The proof of Theorem 4.1 is given in Section 7.5. Note that $\mathcal{U}_{t,k+1} = 0$ when the algorithm uses exact preconditioned gradients at each iteration: $\delta_{t,0,i} = \bar{\mathsf{h}}_i(\widehat{S}_{t,0})$ and $\delta_{t,k+1,i} = \bar{\mathsf{h}}_i(\widehat{S}_{t,k}) - \bar{\mathsf{h}}_i(\widehat{S}_{t,k-1})$ for all $i, t, k$.

**Random number of inner loops $k_t^{\text{in}}$.** When the number of inner loops $k_t^{\text{in}}$ at the outer loop $\#t$ is a random number, we consider it is drawn prior the run of the algorithm. Therefore the expectations in Theorem 4.1 are conditionally to the random sequence $\{k_t^{\text{in}}, t \in [k^{\text{out}}]^\star\}$. The expectation w.r.t. the randomness of $k_t^{\text{in}}$ can easily be obtained from Theorem 4.1; details are omitted.

**The step sizes $\gamma_{t,k}$.** The conditions on the sequence $\{\gamma_{t,k+1}, t \in [k^{\text{out}}]^\star, k \in [k_t^{\text{in}} - 1]\}$ are satisfied with

$$\gamma_{t,k+1} \overset{\text{def}}{=} \prod_{j=0}^{k} \left(1 + \frac{2\,C_b}{m_{t,j+1}}\right)^{-1} \gamma_{t,0}$$

where $\gamma_{t,0}$ is positive and strictly lower than

$$\frac{1}{4Lv_{\max}\upsilon}\frac{\mathsf{b}}{k_t^{\text{in}}}\left(\sqrt{\frac{L_{\dot{\mathrm{W}}}^2}{L^2} + 4v_{\min}v_{\max}\frac{k_t^{\text{in}}}{\mathsf{b}}\upsilon} - \frac{L_{\dot{\mathrm{W}}}}{L}\right) ; \tag{17}$$

$\upsilon \overset{\text{def}}{=} 1 + 2C_{vb}/(\sqrt{\mathsf{b}}\,\inf_{t,k}\bar{M}_{t,k+1})$ (see the proof in Section C.1). First, observe that when $C_b = 0$, the step size can be a constant function of the inner loop index $k$ loop ($\gamma_{t,k+1} = \gamma_{t,0}$ for any $k$). On the contrary, when $C_b > 0$ i.e. for a deterministic approximation or a biased random approximation (see Table 1), the stepsize sequence is a strictly decreasing function of the inner loop index $k$.

Second, the maximal value of $\gamma_{t,0}$ is larger when $C_{vb} = 0$ than when $C_{vb} > 0$. Here again, deterministic or unbiased random approximations requires more aggressive step sizes.

**The initialization of the outer loops.** Set $\mathcal{N}_t \overset{\text{def}}{=} \|\mathcal{E}_t\|^2$. When $\mathcal{B}_{t,0} = \{1, \cdots, n\}$ and $\delta_{t,0,i} = \bar{\mathsf{h}}_i(\widehat{S}_{t,0})$ for all $i$, then $\mathcal{N}_t = 0$; otherwise, $\mathcal{N}_t$ is positive.

Let us discuss the behavior of $\mathcal{N}_t$ when $\delta_{t,0,i}$ is an unbiased random approximation of $\bar{\mathsf{h}}_i(\widehat{S}_{t,0})$ with variance denoted by $\sigma_{t,0,i}^2$. When $\mathcal{B}_{t,0} = \{1, \cdots, n\}$, then

$$\mathbb{E}\left[\mathcal{N}_t\right] = \frac{1}{n^2}\sum_{i=1}^{n}\sigma_{t,0,i}^2 \,. \tag{18}$$

Nevertheless, the strategy $\mathcal{B}_{t,0} = \{1, \cdots, n\}$ has a large computational cost; sampling a subset of size $\mathsf{b}_t'$ reduces the computational cost but increases the squared norm of the error: we have

$$\mathbb{E}\left[\mathcal{N}_t\right] \le \frac{1}{\mathsf{b}_t' n}\sum_{i=1}^{n}\left(\sigma_{t,0,i}^2 + \|\bar{\mathsf{h}}_i(s) - \bar{\mathsf{h}}(s)\|^2\right) \,, \tag{19}$$

with an equality if $\mathcal{B}_{t,0}$ is sampled with replacement in $\{1, \cdots, n\}$. See Section C.2 for detailed computations. From a numerical point of view, an efficient strategy consists in increasing the size $\mathsf{b}_t'$ with the outer loop index $t$ (see references in Section 3.3 for `3P-SPIDER` applied to EM).

**Random stopping time of the algorithm.** In non-convex optimization, the last iterate $\widehat{S}_{k^{\text{out}},k_{k^{\text{out}}}^{\text{in}}}$ is not necessarily the point which minimizes, over the sequence $\{\widehat{S}_{t,k}, t \in [k^{\text{out}}]^\star, k \in [k_t^{\text{in}}]^\star\}$, the distance to the set of solutions of (1). The quantity $\Delta_\star^\star$, motivated by (10), can not be computed exactly in our framework so that the "best" iterate can not be identified thanks to this criterion. It is therefore popular to analyze the algorithm when stopped at a random time (see e.g. (Lan, 2020, Chapter 6)). For sake of simplicity, we consider the case when $k_t^{\text{in}} = k^{\text{in}}$ for any $t$ and $C_b = 0$. We have the following corollary:

**Corollary 4.2** (of Theorem 4.1) *Assume that $k_t^{\text{in}} = k^{\text{in}}$, $C_b = 0$ and the stepsize sequence is constant $\gamma_{t,k} = \gamma_\star$. Let $(\tau, K)$ be a uniform random variable on $[k^{\text{out}}]^\star \times [k^{\text{in}}]^\star$, independent of the algorithm. Then*

$$\begin{aligned}
&\inf_{(t,k)\in[k^{\text{out}}]^\star\times[k^{\text{in}}]^\star}\left(\frac{1}{2} - \Lambda_{t,k}\right)\mathbb{E}\left[\Delta_{\tau,K}^\star + \mathcal{D}_{\tau,K}^\star\right] \\
&\le \frac{\mathbb{E}\left[\mathrm{W}(\widehat{S}_{1,0}) + g(\widehat{S}_{1,0})\right] - \min_{\mathcal{S}}\left(\mathrm{W}+g\right)}{k^{\text{out}}k^{\text{in}}\gamma_\star} \\
&+ v_{\max}\mathbb{E}\left[\|\mathcal{E}_\tau\|^2\right] + v_{\max}\,\mathbb{E}\left[\left(k^{\text{in}} - K + 1\right)\mathcal{U}_{\tau,K}\right] \,.
\end{aligned}$$

An upper bound on $\Lambda_{t,k}$ can easily be obtained from (15) as a function of $L_{\dot{\mathrm{W}}}, L, v_{\min}, v_{\max}$, $k^{\mathrm{in}}, \mathsf{b}, \gamma_\star$, $C_{vb}$ and $\inf_{t,k} \bar{M}_{t,k+1}$.

Corollary 4.2 shows that, even by stopping `3P-SPIDER` with this simple rule, the first term in the RHS is inversely proportional to the maximal number of iterations $k^{\mathrm{out}} k^{\mathrm{in}}$.

**Complexity analysis when $\mathcal{E}_t = 0$, $\mathcal{U}_{t,k} = 0$ and $k_t^{\mathrm{in}} = k^{\mathrm{in}}$.** For smooth first-order optimization, algorithms are compared through their complexity in order to satisfy an $\epsilon$-first order stationary condition. In stochastic composite optimization, this criterion is naturally extended to the *$\epsilon$-approximate stationary condition* defined by

$$\mathbb{E}\left[\Delta^\star_{\tau,K}\right] \leq \epsilon \,,$$

where $(\tau, K)$ is a random variable taking values in $[k^{\mathrm{out}}]^\star \times [k^{\mathrm{in}}]^\star$; see e.g. (Ghadimi et al, 2016, Section 4), (Wang et al, 2019, Section 3) and Fort and Moulines (2021)).

Corollary 4.3 studies the proximal complexity $\mathcal{K}_{\mathrm{prox}}$ defined as the number of calls to the prox operator in order to satisfy the $\epsilon$-approximate stationary condition; the stochastic $\bar{\mathsf{h}}$-complexity $\mathcal{K}_{\bar{\mathsf{h}}}$ defined as the number of calls to one of the $\bar{\mathsf{h}}_i$'s; and the total number of iterations $k^{\mathrm{in}} k^{\mathrm{out}}$. Again for sake of simplicity, and in order to compare our results to the literature, we consider a simplified setting.

**Corollary 4.3** (of Corollary 4.2) *Assume in addition that $\mathcal{E}_t = 0$ and $\mathcal{U}_{t,k} = 0$. The $\epsilon$-approximate stationary condition is satisfied with $\gamma_\star = v_{\min}/(4\, L_{\dot{\mathrm{W}}})$, $k^{\mathrm{in}}/\mathsf{b} = L_{\dot{\mathrm{W}}}^2/(v_{\min} v_{\max} L^2)$, $\mathsf{b} = O(\sqrt{n}\sqrt{v_{\min} v_{\max}} L / L_{\dot{\mathrm{W}}})$ and $k^{\mathrm{out}} k^{\mathrm{in}} = O(L_{\dot{\mathrm{W}}}/(\epsilon v_{\min}))$. Moreover, $\mathcal{K}_{\mathrm{prox}} = O(L_{\dot{\mathrm{W}}}/(v_{\min}\, \epsilon))$ and $\mathcal{K}_{\bar{\mathsf{h}}} = O(\sqrt{v_{\max}} L \sqrt{n}/(\epsilon \sqrt{v_{\min}}))$.*

The proof is in Section 7.6. This result shows that the step size $\gamma_\star$ and the number of inner loops $k^{\mathrm{in}}$ are independent of the accuracy $\epsilon$.

When applied to Stochastic Gradient Descent, `3P-SPIDER` in the setting of Corollary 4.3 is the `Prox-SpiderBoost` algorithm studied in Wang et al (2019): Corollary 4.3 and (Wang et al, 2019, Theorem 2) state the same complexity results. (Wang et al, 2019, Table 1) compares `Prox-SpiderBoost` to other stochastic gradient algorithms for composite non-convex finite sum optimization. It is shown that the variance reduction based on `SPIDER` order-level outperforms other variance reduction strategies such as the `SVRG` one and the `SAGA` on, introduced respectively by Johnson and Zhang (2013) and Defazio et al (2014). Hence, `3P-SPIDER` reaches the state of the art among the proximal stochastic gradient algorithms designed to solve finite sum non-convex composite optimization.

When applied to EM, `3P-SPIDER` in the setting of Corollary 4.3 is the extension of the `SPIDER-EM` algorithm studied in Fort et al (2020) to the case there is a proximal step which manages the constraint $g$. Here again, the comparison of Corollary 4.3 and (Fort et al, 2020, Theorem 2) shows that `3P-SPIDER` reaches the state of the art among the incremental EM algorithms with variance reduction, including `sEM-VR` and `FIEM` introduced respectively in Chen et al (2018) and Karimi et al (2019) (see also Fort et al (2021a). See the comparison to the literature in Fort et al (2020).

Beyond these two applications, Corollary 4.3 is - to our best knowledge - the first complexity result for an algorithm designed to solve (1) under the constraint (2) and for non-convex finite sum composite optimization.

**$\epsilon$-approximate stationary condition: the cost of inexact preconditioned forward operators.** Let us discuss the cost of inexact $\bar{\mathsf{h}}_i(s)$'s when the approximation is unbiased and random (so that $C_b = C_{vb} = 0$, see Table 1): does it deteriorate the proximal complexity $\mathcal{K}_{\mathrm{prox}}$ and the number of calls to an oracle of a preconditioned forward operator $\bar{\mathsf{h}}_i$ (still denoted by $\mathcal{K}_{\bar{\mathsf{h}}}$ below) ? detailed computations of the assertions below can be found in Section 7.7.

If $\mathbb{E}\left[\|\mathcal{E}_t\|^2\right] = O(\epsilon^{1-\mathsf{a}'}/(\sqrt{n}t)^{\mathsf{a}'})$ for some $\mathsf{a}' \in [0,1)$ and

$$M_{t,k+1} = O\left(\frac{n^{(\mathsf{a}-\bar{\mathsf{a}})/2}}{\epsilon^{1-\mathsf{a}}}\, t^{\mathsf{a}}\, (k+1)^{\bar{\mathsf{a}}}\right)$$

for some $\mathsf{a}, \bar{\mathsf{a}} \in [0,1)$, then the $\epsilon$-approximate stationary condition is satisfied with $k_t^{\mathrm{in}} = O(\sqrt{n})$, $\mathsf{b} = O(\sqrt{n})$ and $k^{\mathrm{out}} = O(1/(\sqrt{n}\epsilon))$. In addition, $\mathcal{K}_{\mathrm{prox}} = O(1/\epsilon)$ and $\mathcal{K}_{\bar{\mathsf{h}}} = O(\sqrt{n}/\epsilon)$. Therefore, the conclusions of Corollary 4.3 remain valid, and the approximations of the $\mathsf{h}_i$'s do not deteriorate

the complexity performances of the algorithms, as soon as the approximation is small enough.

Let us now evaluate the computational cost, in the case the unbiased random approximation is a Monte Carlo approximation computed from independent and identically distributed (i.i.d.) samples. In this case, $M_{t,\cdot}$ is the number of terms of the Monte Carlo sum (see Section A). The Monte Carlo complexity $\mathcal{K}_{\mathsf{MC}}$ defined as the total number of Monte Carlo draws required to satisfy the $\epsilon$-approximate stationary condition is: $\mathcal{K}_{\mathsf{MC}} = O(\sqrt{n}/\epsilon^2)$ for any $\mathsf{a}, \mathsf{a}', \bar{\mathsf{a}} \in [0, 1)$.

To our best knowledge, it is the first complexity analysis with such a Monte Carlo approximation of the preconditioned forward operators $\bar{\mathsf{h}}_i$'s.

# 5 Application: Penalized Logistic Regression with random effects

## 5.1 The model

Motivated by applications in classification, we consider a logistic regression model with random effects.

Let $n$ pairs of examples $\{(X_i, Y_i), i \in [n]^\star\}$ where $X_i \in \mathbb{R}^d$ collects the $d$ explanatory variables, and $Y_i$ is the binary response variable taking values in $\{-1, 1\}$. We assume that given $\{X_i, i \in [n]^\star\}$, the binary observations $\{Y_i, i \in [n]^\star\}$ are independent with distribution

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

In words, each example $\#i$ has an individual regression vector $Z_i$ in $\mathbb{R}^d$ and given $Z_i$, the success probability $\mathbb{P}(Y_i = 1 \mid Z_i)$ is $(1+\exp(-\langle X_i, Z_i\rangle))^{-1}$. The regression vectors $Z_1, \cdots, Z_n$ are independent with a Gaussian distribution $\mathcal{N}(\theta, \sigma^2 \mathrm{I}_d)$. $\theta$ is assumed to be unknown and $\sigma^2$ is known.

The objective is the estimation of $\theta$ by maximizing the penalized log-likelihood criterion, with a ridge penalty $\mathrm{pen}(\theta) \stackrel{\mathrm{def}}{=} n\tau\|\theta\|^2$, where $\tau > 0$. By a change of variable, we obtain that the criterion to be minimized is (see Lemma B.1)

$$F : \theta \mapsto -\frac{1}{n}\sum_{i=1}^n \log \int_{\mathbb{R}} \frac{\exp\left(x \langle X_i, \theta\rangle / (\sigma^2 \|X_i\|)\right)}{1+\exp(-y_i\|X_i\|x)} \times \exp\left(-x^2/(2\sigma^2)\right) \mathrm{d}x + \|\theta\|_U^2 \ ,$$

where

$$U \stackrel{\mathrm{def}}{=} \tau \mathrm{I}_d + \frac{1}{2\sigma^2}\frac{1}{n}\sum_{i=1}^n \frac{X_i X_i^\top}{\|X_i\|^2} \ .$$

The following lemma shows that the minimizers of $F$ are in a compact set $\mathcal{K}$ of $\mathbb{R}^d$ thus implying that the optimization problem can be constrained to $\mathcal{K}$. The proof is given in Section B.2.

**Lemma 5.1** *The minimizers of $F$ are in the set $\mathcal{K} \stackrel{\mathrm{def}}{=} \{\theta \in \mathbb{R}^d : \|\theta\|^2 \le (\ln 4)/\tau\}$.*

To solve this optimization problem, we propose two approaches: a gradient one, solved in the original space $\theta \in \mathbb{R}^d$ (see Section 2.1); and an EM one, solved in the statistic space (see Section 2.2). The discussions in Section 5.2 and Section 5.3 show that EM is a gradient approach for finding the critical points of $s \mapsto F(U^{-1}s/2)$.

## 5.2 A Gradient approach

We are interested in finding a critical point of $F$ in $\mathcal{K}$. Equivalently, we want to solve

$$0 \in \frac{1}{n}\sum_{i=1}^n G_i(\theta) + \partial g(\theta)$$

where $g$ is the $\{0, +\infty\}$-valued indicator function of the set $\mathcal{K}$ and

$$G_i(\theta) \stackrel{\mathrm{def}}{=} 2U\theta - \frac{X_i}{\sigma^2\|X_i\|}\int_{\mathbb{R}} z \ \pi_{\theta,i}(z)\mathrm{d}z \ ;$$

$\pi_{\theta,i}(z)$ is the probability density proportional to

$$\frac{\exp\left(z \langle X_i, \theta\rangle / (\sigma^2\|X_i\|) - z^2/(2\sigma^2)\right)}{1+\exp(-y_i\|X_i\|z)} \ . \qquad (20)$$

We apply `3P-SPIDER` with $B \stackrel{\mathrm{def}}{=} \mathrm{I}_q$ and $\mathsf{h}_i(\theta, \mathrm{I}_q) \stackrel{\mathrm{def}}{=} -G_i(\theta)$; note that $\mathrm{prox}_{\gamma g}(\theta) = \mathrm{argmin}_{x\in\mathcal{K}}\|x -$

$\theta\|^2$. $\mathsf{h}_i$ is the sum of an explicit term and an integral with no closed form: it will be approximated by a Monte Carlo method, based on a Markov chain Monte Carlo (MCMC) sampler (see Section 5.4 below). Therefore, $\delta_{t,k,i}$ will be a biased random approximation.

## 5.3 An EM approach

The criterion $F$ to be minimized is of the form (5) with $\mathsf{Z} = \mathbb{R}$, $\mu_{lv}(\mathrm{d}z) = \mathrm{d}z$ and $p(Y_i, z;\ \theta)$ equal to

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

The curved exponential family assumption on the complete data model is satisfied: $p(Y_i, z;\ \theta) = H(Y_i, z)\,\exp\left(\langle S(Y_i, z), \phi(\theta)\rangle - \psi(\theta)\right)$ with $\phi(\theta) \stackrel{\text{def}}{=} \theta$, $\psi(\theta) \stackrel{\text{def}}{=} \|\theta\|_U^2$ and

$$S(Y_i, z) \stackrel{\text{def}}{=} z\,\frac{X_i}{\sigma^2\,\|X_i\|}\ .$$

From Section 2.2, EM in the statistic space is of the form (1)-(2): it solves $0 \in n^{-1}\sum_{i=1}^n \bar{G}_i(s) + \partial\bar{g}(s)$ where $\bar{G}_i(s) = B\,G_i(Bs)$, $B \stackrel{\text{def}}{=} U^{-1}/2$ and $\bar{g}(s)$ is the $\{0, +\infty\}$-valued indicator function of the set $\{s \in \mathbb{R}^d : \mathsf{T}(s) \in \mathcal{K}\}$ where $\mathsf{T}(s) \stackrel{\text{def}}{=} Bs$; it uses

$$\bar{\mathsf{h}}_i(s) \stackrel{\text{def}}{=} \frac{X_i}{\sigma^2\,\|X_i\|}\int_{\mathbb{R}} z\,\pi_{Bs,i}(z)\mathrm{d}z - s\ , \qquad (21)$$

and the metric induced by $\mathsf{B}(s) \stackrel{\text{def}}{=} B$. See Section B.3 for detailed computations. As in the gradient approach, $\mathsf{h}_i$ requires the expectation of the distribution $\pi_{\cdot,i}$ (see (20)) which has no closed form. We will run `3P-SPIDER` with $B(s) \leftarrow B$ and a biased random approximation of the $\bar{\mathsf{h}}_i(s)$'s (see Section 5.4); note that $\mathrm{prox}^B_{\gamma\,\bar{g}}(s) = B^{-1}\,\mathrm{argmin}_{x\in\mathcal{K}}\left((x-Bs)^\top B^{-1}(x-Bs)\right)$.

## 5.4 The MCMC approximation of $\bar{\mathsf{h}}_i$

We discuss how to design an efficient MCMC sampler for the approximation of

$$\mathcal{I}_i(\theta) \stackrel{\text{def}}{=} \int_{\mathbb{R}} z\,\pi_{\theta,i}(z)\,\mathrm{d}z\ , \quad \theta \in \mathbb{R}^d\ ,$$

where $\pi_{\theta,i}$ is defined, up to a normalizing constant, by (20). By using an integration by parts and by applying (Polson et al, 2013, Theorem 1), we show that a data augmentation scheme is possible to approximate integrals w.r.t. $\pi_{\theta,i}(z)$.

**Lemma 5.2** *For any $i \in [n]^\star$ and $\theta \in \mathbb{R}^d$, it holds*

$$\mathcal{I}_i(\theta) = \left\langle \frac{X_i}{\|X_i\|}, \theta \right\rangle + y_i\|X_i\|\sigma^2\int_{\mathbb{R}}\int_0^{+\infty} \frac{\bar{\pi}_{\theta,i}(z,\omega)}{1+\exp\left(y_i\|X_i\|z\right)}\,\mathrm{d}z\mathrm{d}\omega\ ,$$

*where $\bar{\pi}_{\theta,i}(z,\omega)$ is a probability density on $\mathbb{R}\times(0,+\infty)$. The conditional distribution of $z$ given $\omega$ is a Gaussian distribution with parameters*

$$\frac{\langle X_i,\theta\rangle\,/\|X_i\| + y_i\|X_i\|\sigma^2/2}{1+\omega\sigma^2\|X_i\|^2}\ , \qquad \frac{\sigma^2}{1+\omega\sigma^2\|X_i\|^2}\ ;$$

*the conditional distribution of $\omega$ given $z$ is a Polya-Gamma distribution with parameters $(1, \|X_i\|z)$.*

The proof is given in Section B.4. Therefore, a Monte Carlo approximation of integrals w.r.t. $\pi_{\theta,i}$ are obtained from a Gibbs sampler targeting the distribution $\bar{\pi}_{\theta,i}(z,\omega)$: it produces a sequence of pairs $\{(Z_r, \Omega_r), r \geq 0\}$ and only the $Z_r$'s are retained for the Monte Carlo approximation. For example, $\bar{\mathsf{h}}_i(s)$ given by (21) can be approximated by

$$\bar{\mathsf{h}}_i(s) \approx -s + \frac{X_i}{\sigma^2\|X_i\|^2}\,\langle X_i, Bs\rangle + y_i\|X_i\|\frac{1}{m}\sum_{r=1}^m \left(1+\exp(y_i\|X_i\|Z_r^{s,i})\right)^{-1}\ . \qquad (22)$$

This Gibbs sampler is uniformly ergodic (see (Choi and Hobert, 2013, Proposition 3.1)); consequently, upon noting that $z \mapsto H_i(z) \stackrel{\text{def}}{=} (1+\exp(y_i\|X_i\|z))^{-1}$ is bounded by one uniformly in $i$ and $z$, the conditions A5 in Section A are verified with $U$ equal to the constant function 1 and with a geometric convergence rate $\rho(r) \stackrel{\text{def}}{=} \upsilon^r$ for some $\upsilon \in (0,1)$ (remember that $\mathcal{S}$ is a compact set in our application); details are provided in Section B.5.

Therefore, A4 is verified and the rates $m_{t,k+1}$, $M_{t,k+1}$ and $\bar{M}_{t,k+1}$ are equal, and equal to the number of points in the Monte Carlo sum (see Proposition A.1).

## 5.5 Numerical illustrations

Let us run 3P-SPIDER for minimizing the criterion $F$; based on previous results comparing variance reduced Expectation Maximization algorithms and variance reduced Gradient algorithms (see e.g. (Chen et al, 2018, section 4)), we restrict our attention to the EM approach. In this numerical application, $n = 24\,989$ and $d = 21$; we choose $\tau = 1$ and $\sigma^2 = 0.05$.

**The data set.** The $n$ pairs $(y_i, X_i)$ are built from the MNIST data set. The 13 007 examples labeled $y_i = -1$ are the examples labeled 1 or 7 in the MNIST training data set; the 11 982 examples labeled $y_i = 1$ are the examples labeled 3 or 8 in the MNIST training data set. The covariates $X_i$ are obtained as follows. Let $\mathsf{X}^{\mathrm{im}}$ be the $784 \times n$ matrix collecting the 784 pixels for each image. The pixels take values in $[0, 1]$. Then the rows of $\mathsf{X}^{\mathrm{im}}$ are centered; by a PCA, each image is reduced to a vector in $\mathbb{R}^{20}$. This yields $\mathsf{X}^{\mathrm{red}} \in \mathbb{R}^{20\times n}$. Finally, $\mathsf{X}^{\mathrm{red}}$ is augmented with a row of ones, yielding $\mathsf{X} \in \mathbb{R}^{21\times n}$. The columns of $\mathsf{X}$ are the $X_i$'s.

**The algorithms.** We compare four algorithms. EM denotes the SAEM algorithm (Delyon et al (1999)) combined with a proximal step: each iteration processes the full data set so that there is one iteration of EM per epoch:

$$\widehat{S}^{\mathtt{EM}}_{r+1} \stackrel{\mathrm{def}}{=} \mathrm{prox}^B_{\gamma g}(\widehat{S}^{\mathtt{EM}}_r + \frac{\gamma}{n}\sum_{i=1}^n \widehat{\bar{\mathsf{h}}_i(\widehat{S}^{\mathtt{EM}}_r)}) \; .$$

Online EM is the algorithm given by Cappé and Moulines (2009) combined with a proximal step; each iteration processes $\mathsf{b}$ examples and below, we will run $k^{\mathrm{in}} \stackrel{\mathrm{def}}{=} \lceil n/\mathsf{b} \rceil$ iterations per epoch:

$$\widehat{S}^{\mathtt{OEM}}_{r+1} \stackrel{\mathrm{def}}{=} \mathrm{prox}^B_{\gamma g}(\widehat{S}^{\mathtt{OEM}}_r + \frac{\gamma}{\mathsf{b}}\sum_{i\in\mathcal{B}_{r+1}} \widehat{\bar{\mathsf{h}}_i(\widehat{S}^{\mathtt{OEM}}_r)}) \; .$$

For EM and Online EM, $\widehat{\bar{\mathsf{h}}_i(\widehat{S}^{\bullet}_t)}$ is a Monte Carlo approximation of $\bar{\mathsf{h}}_i(\widehat{S}^{\bullet}_t)$ computed with $m^t$ points. 3P-SPIDER is Algorithm 2; we choose $k^{\mathrm{in}}_t = k^{\mathrm{in}}$ and $k^{\mathrm{in}} = \lceil n/\mathsf{b} \rceil$ so that one epoch corresponds to the $k^{\mathrm{in}}$ inner loops; we choose $\mathsf{b}'_t = n$ so that the initialization of each outer loop is one epoch; the $\delta_{t,k,i}$ are computed by Monte Carlo sums (see (22)) with $m^0$ points for $\delta_{t,0,i}$ and $m^t$ points for $\delta_{t,k+1,i}$; since $\widehat{S}_{t,0} = \widehat{S}_{t,-1}$, we set $\delta_{t,1,i} = 0$ for all $i$, so that $\mathsf{S}_{t,1} = \mathsf{S}_{t,0} = n^{-1}\sum_{i=1}^n \delta_{t,0,i}$. Finally, 3P-SPIDER and 3P-SPIDER-corr differ as follows: the Monte Carlo approximation $\delta_{t,k+1,i}$ necessitates a Monte Carlo approximation of $\bar{\mathsf{h}}_i(\widehat{S}_{t,k})$ and one of $\bar{\mathsf{h}}_i(\widehat{S}_{t,k-1})$. In 3P-SPIDER, the Monte Carlo approximations are based on two independent chains (see (22)) while in 3P-SPIDER-corr the chains are correlated.

All the algorithms are initialized at the null vector $\widehat{S}_{\mathrm{init}} = 0 \in \mathbb{R}^d$. The step size is equal to $\gamma = 0.4$ during the first six epochs and then equal to $\gamma = 0.1$. The length of all the paths is 20 epochs. On all the figures except Figure 3, we report a mean value computed over 25 independent runs of each algorithm; the shadowed area is delimited by the minimal and maximal value of the displayed criterion over these runs.

**Analyses.** Most of the comparisons are based on the evolution of

$$\Delta_{t,k+1} \stackrel{\mathrm{def}}{=} \frac{\|\mathrm{prox}^B_{\gamma g}(\widehat{S}_{t,k} + \gamma\,\mathsf{S}_{t,k+1}) - \widehat{S}_{t,k}\|^2_B}{\gamma^2}$$

as a function of the number of epochs; this criterion is an approximation of $\Delta^\star_{t,k+1}$ (see (14)) which can not be computed here since $\bar{\mathsf{h}}$ has no closed form in this application. The criterion $\Delta_{t,k+1}$ for 3P-SPIDER and 3P-SPIDER-corr, is compared to $\Delta^{\mathtt{EM}}_r$ defined by

$$\frac{\|\mathrm{prox}^B_{\gamma g}(\widehat{S}^{\mathtt{EM}}_r + \gamma n^{-1}\sum_{i=1}^n \widehat{\bar{\mathsf{h}}_i(\widehat{S}^{\mathtt{EM}}_r)}) - \widehat{S}^{\mathtt{EM}}_r\|^2_B}{\gamma^2} \; ;$$

and to $\Delta^{\mathtt{OEM}}_r$ defined by

$$\frac{\|\mathrm{prox}^B_{\gamma g}(\widehat{S}^{\mathtt{OEM}}_r + \gamma \mathsf{b}^{-1}\sum_{i\in\mathcal{B}_{r+1}} \widehat{\bar{\mathsf{h}}_i(\widehat{S}^{\mathtt{OEM}}_r)}) - \widehat{S}^{\mathtt{OEM}}_r\|^2_B}{\gamma^2} \; .$$

The best algorithm will have the smallest value of $\Delta_{t,k+1}$.

We first study the role of some design parameters of 3P-SPIDER, such as the number of Monte Carlo points when computing $\delta_{t,0,i}$ (denoted by $m^0$) and $\delta_{t,k+1,i}$ (denoted by $m^t$) and the balance between $k^{\mathrm{in}}$ and $\mathsf{b}$ which satisfy $k^{\mathrm{in}}\mathsf{b} \approx n$.

On Figure 1, two strategies are chosen: first, $m^0 = m^t = 2\lceil\sqrt{n}\rceil$; then $m^0 = m^t = 5\lceil\sqrt{n}\rceil$; in all cases, $k^{\mathrm{in}} = \lceil\sqrt{n}/10\rceil$ and $\mathsf{b} = \lceil n/k^{\mathrm{in}}\rceil$. For comparison, EM and Online EM are also run, with a number of Monte Carlo point equal to $m^t$ at each iteration.

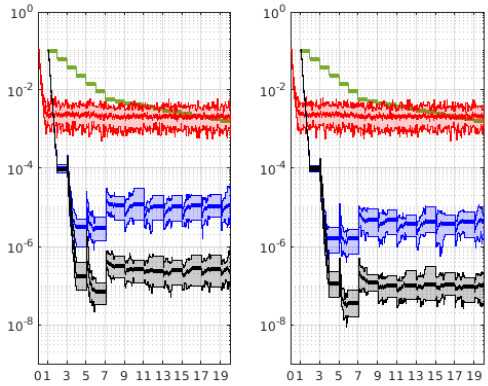

**Fig. 1** Different strategies for the number of Monte Carlo points when approximating $\bar{\mathsf{h}}_i(s)$ - see (22). Evolution of $\Delta_r^{\texttt{EM}}$ in green, $\Delta_r^{\texttt{OEM}}$ in red, $\Delta_{t,k+1}$ for `3P-SPIDER` in blue and $\Delta_{t,k+1}$ for `3P-SPIDER corr` in black, as a function of the number of epochs. [left] $m^0 = m^t = 2\lceil\sqrt{n}\rceil$, [right] $m^0 = m^t = 5\lceil\sqrt{n}\rceil$.

On Figure 2, the case when $k^{\text{in}} = \lceil\sqrt{n}/10\rceil$ is compared to the case $k^{\text{in}} = \lceil\sqrt{n}/2\rceil$; in both cases, $\mathsf{b} = \lceil n/k^{\text{in}}\rceil$ and $m^0 = m^t = 2\lceil\sqrt{n}\rceil$.

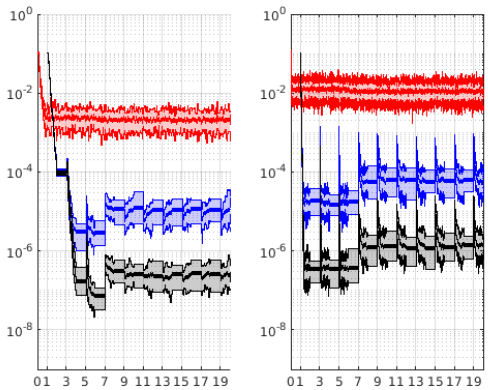

**Fig. 2** Number of inner loops per epoch. Evolution of $\Delta_r^{\texttt{OEM}}$ in red, $\Delta_{t,k+1}$ for `3P-SPIDER` in blue and $\Delta_{t,k+1}$ for `3P-SPIDER corr` in black, as a function of the number of epochs. [left] $k^{\text{in}} = \lceil\sqrt{n}/10\rceil$ and $\mathsf{b} = \lceil n/k^{\text{in}}\rceil$. [right] $k^{\text{in}} = \lceil\sqrt{n}/2\rceil$ and $\mathsf{b} = \lceil n/k^{\text{in}}\rceil$.

Each algorithm returns a sequence of points in the $s$-space, from which a sequence of points in the $\theta$-space is deduced through the formula $\theta = \mathsf{T}(s) = Bs \in \mathbb{R}^d$. On Figure 3, three components of this $\theta$-sequence are displayed, versus the number of epochs.

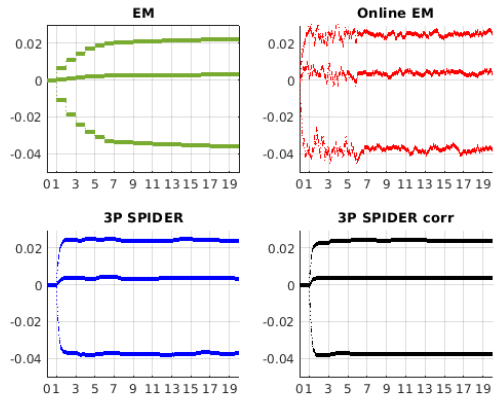


**Fig. 3** Estimation of three parameters. Evolution of the three components of $\theta$ by `EM` in green (top, left), `OEM` in red (top, right), `3P-SPIDER` in blue (bottom, left) and `3P-SPIDER corr` in black (bottom, right), as a function of the number of epochs.

Finally, we also display on Figure 4 the evolution of the squared norm of the iterates $\|\widehat{S}_{t,k}\|^2$ obtained by `3P-SPIDER` and `3P-SPIDER-corr`, and $\|\widehat{S}_r^{\texttt{OM}}\|^2$ and $\|\widehat{S}_r^{\texttt{OEM}}\|^2$ obtained resp. by `EM` and `Online EM`. They are plotted as a function of the epochs.

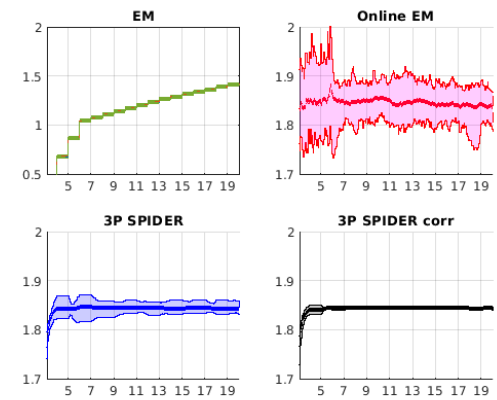


**Fig. 4** Squared norm of the iterates. Evolution of $\|\widehat{S}_r^{\texttt{EM}}\|^2$ in green (top, left), $\|\widehat{S}_r^{\texttt{OEM}}\|^2$ in red (top, right), $\|\widehat{S}_{t,k}\|^2$ for `3P-SPIDER` in blue (bottom, left) and $\|\widehat{S}_{t,k}\|^2$ for `3P-SPIDER corr` in black (bottom, right), as a function of the number of epochs.

**Conclusions.** `EM` has a slow convergence rate and even fails to converge before 20 epochs contrary to the other algorithms (see e.g. Figure 4): one update of the iterate per epoch is not enough especially during the first iterations when more

updates even based on part of the data set is a better strategy (see e.g. the behavior of `Online EM`, which contains $k^{\text{in}}$ updates per epoch).
`Online EM`, `3P-SPIDER` and `3P-SPIDER-corr` process part of the data set at each iteration; compared to `Online EM`, the `3P-SPIDER`'s contain a variance reduction. All the plots illustrate the benefit of this variance reduction, which reduces the variability at convergence.
The choice of $\gamma$ impacts this variability: see e.g. Figures 1, 2 and Figure 4 where a change occurs at epoch #7 (remember that from epoch $2\ell$ to $2\ell+1$, `Online EM` runs $k^{\text{in}}$ updates of the iterates while the `3P-SPIDER`'s do not update the iterate since they compute $\mathsf{S}_{t,0}$).
`3P-SPIDER-corr` improves on `3P-SPIDER`. The control variate has a larger impact when the correlation is increased, as illustrated by all plots. It decreases the variability introduced by the mini-batches ($\mathsf{b} < n$) and the variability introduced by the Monte Carlo approximation $\delta_{t,k+1,i}$.
Given the budget of $n$ examples processed per outer loops, Figure 2 shows that at convergence, the accuracy is improved by larger mini batch sizes and therefore a smaller number of inner loops. Not surprisingly, a larger number of Monte Carlo points decreases the variability at convergence (see Figure 1).

# 6 Proof of Section 3

## 6.1 Proof of Lemma 3.1

Lemma 6.1 collects the two statements of Lemma 3.1 and a third property.

**Lemma 6.1** *Assume* A*1.*

1. *For any $\gamma > 0$, $B \in \mathcal{P}_+^q$ and $s \in \mathbb{R}^q$, the optimization problem (7) has a unique solution, characterized as the unique point $\mathsf{p} \in \mathcal{S}$ satisfying $-\gamma^{-1} B(\mathsf{p} - s) \in \partial g(\mathsf{p})$.*
2. *For any $\gamma > 0$, $B \in \mathcal{P}_+^q$, $s \in \mathcal{S}$ and $h \in \mathbb{R}^q$,*

$$s = \operatorname{prox}_{\gamma g}^B(s + \gamma h) \quad \textit{iff} \quad Bh \in \partial g(s). \tag{23}$$

3. *Let $\gamma > 0$ and $B \in \mathcal{P}_+^q$. The operator $\operatorname{prox}_{\gamma g}^B$ is firmly nonexpansive; this implies that for any $s, s' \in \mathbb{R}^q$,*

$$\|\operatorname{prox}_{\gamma g}^B(s') - \operatorname{prox}_{\gamma g}^B(s)\|_B^2 \leq \left\langle \operatorname{prox}_{\gamma g}^B(s') - \operatorname{prox}_{\gamma g}^B(s), s' - s \right\rangle_B .$$

*Proof* Existence, uniqueness and characterization are established in Hiriart-Urruty and Lemaréchal (1996, Chapter XV, Lemma 4.1.1). The statement (23) follows from the characterization; note that $\operatorname{prox}_{\gamma g}^B(s) \in \mathcal{S}$ for any $s \in \mathbb{R}^q$. The firmly nonexpansive property is a consequence of Hiriart-Urruty and Lemaréchal (1996, Chapter XV, Theorem 4.1.4). □

# 7 Proof of Section 4

## 7.1 Notations

Define for any $s \in \mathcal{S}$,

$$\bar{\mathsf{h}}(s) \stackrel{\text{def}}{=} \frac{1}{n} \sum_{i=1}^n \bar{\mathsf{h}}_i(s) , \qquad \bar{\mathsf{h}}_{\mathcal{B}} \stackrel{\text{def}}{=} \frac{1}{\mathsf{b}} \sum_{i \in \mathcal{B}} \bar{\mathsf{h}}_i ,$$

where $\mathcal{B}$ is an $n$-tuple of elements of $[n]^\star$ (with or without multiplicity) of cardinal $\mathsf{b}$.

All the random variables are defined on a probability space $(\Omega, \mathcal{A}, \mathbb{P})$. It is endowed with the following filtrations for $t \geq 0$ and $k \geq 0$,

$$\begin{aligned}
\mathcal{F}_{0,k_0^{\text{in}}} &\stackrel{\text{def}}{=} \sigma(\widehat{S}_{\text{init}}), \\
\mathcal{F}_{t,0} &\stackrel{\text{def}}{=} \mathcal{F}_{t-1,k_{t-1}^{\text{in}}} \bigvee \sigma\left(\mathcal{B}_{t,0}, \delta_{t,0,i} \text{ for all } i\right), \\
\mathcal{F}_{t,k+\frac{1}{2}} &\stackrel{\text{def}}{=} \mathcal{F}_{t,k} \bigvee \sigma(\mathcal{B}_{t,k+1}), \\
\mathcal{F}_{t,k+1} &\stackrel{\text{def}}{=} \mathcal{F}_{t,k+\frac{1}{2}} \bigvee \sigma\left(\delta_{t,k+1,i} \text{ for all } i \in \mathcal{B}_{t,k+1}\right) .
\end{aligned}$$

For any $t \in [k^{\text{out}}]^\star$, set

$$\mathcal{E}_t \stackrel{\text{def}}{=} \mathsf{S}_{t,0} - \bar{\mathsf{h}}(\widehat{S}_{t,0}) = \frac{1}{\mathsf{b}'_t} \sum_{i \in \mathcal{B}_{t,0}} \delta_{t,0,i} - \bar{\mathsf{h}}(\widehat{S}_{t,0}) .$$

$\mathcal{E}_t$ is the error when replacing the full sum using exact terms $\bar{\mathsf{h}}_i(\widehat{S}_{t,0})$, with a possibly subsum of size $\mathsf{b}'_t < n$ using approximations of $\bar{\mathsf{h}}_i(\widehat{S}_{t,0})$. Remember that

$$\xi_{t,k+1,i} \stackrel{\text{def}}{=} \delta_{t,k+1,i} - \bar{\mathsf{h}}_i(\widehat{S}_{t,k}) + \bar{\mathsf{h}}_i(\widehat{S}_{t,k-1}) ,$$

and

$$\begin{aligned}
\mu_{t,k+1,i} &\stackrel{\text{def}}{=} \mathbb{E}\left[\xi_{t,k+1,i} | \mathcal{F}_{t,k+1/2}\right] , \\
\sigma^2_{t,k+1,i} &\stackrel{\text{def}}{=} \mathbb{E}\left[\|\xi_{t,k+1,i} - \mu_{t,k+1,i}\|^2 | \mathcal{F}_{t,k+1/2}\right] .
\end{aligned}$$

Finally, set

$$\eta_{t,k+1} \stackrel{\text{def}}{=} \frac{1}{\mathsf{b}} \sum_{i \in \mathcal{B}_{t,k+1}} \xi_{t,k+1,i}.$$

Throughout the proof, we will use the shorthand notation

$$B_{t,k} \stackrel{\text{def}}{=} \mathsf{B}(\widehat{S}_{t,k}) \, .$$

## 7.2 Preliminary lemmas

**Lemma 7.1** *Let $\mathcal{B}$ be a batch of $[n]^\star$ of size $\mathsf{b}$, sampled at random (with or without replacement).*

1. *For any family $\{f_1, \cdots, f_n\}$, $\mathbb{E}\left[\mathsf{b}^{-1} \sum_{i \in \mathcal{B}} f_i\right] = n^{-1} \sum_{i=1}^n f_i$.*
2. *For any family $\{f_1, \cdots, f_n\}$,*

$$\mathbb{E}\left[\left\| \frac{1}{\mathsf{b}} \sum_{i \in \mathcal{B}} f_i - \frac{1}{n} \sum_{i=1}^n f_i \right\|^2\right] \leq \frac{1}{\mathsf{b}\, n} \sum_{i=1}^n \| f_i - \frac{1}{n} \sum_{j=1}^n f_j \|^2 \, .$$

3. *Assume A2. For any $s, s' \in \mathcal{S}$, it holds*

$$\mathbb{E}\left[\left\| \left\{\bar{\mathsf{h}}_{\mathcal{B}}(s) - \bar{\mathsf{h}}_{\mathcal{B}}(s')\right\} - \left\{\bar{\mathsf{h}}(s) - \bar{\mathsf{h}}(s')\right\} \right\|^2\right] \leq \frac{1}{\mathsf{b}} \left( L^2 \|s - s'\|^2 - \|\bar{\mathsf{h}}(s) - \bar{\mathsf{h}}(s')\|^2 \right) \, .$$

*Proof* The proof is along the same lines as the proof of (Fort et al, 2020, Lemma 4). A detailed proof is provided in Section C.3. □

**Lemma 7.2** *Assume A4-item 1 and A4-item 3. For any $t \in [k^{\text{out}}]^\star$ and $k \in [k_t^{\text{in}} - 1]$, it holds*

$$\mathbb{E}\left[\eta_{t,k+1} | \mathcal{F}_{t,k+1/2}\right] = \frac{1}{\mathsf{b}} \sum_{i \in \mathcal{B}_{t,k+1}} \mu_{t,k+1,i} \, ,$$

$$\mathbb{E}\left[\eta_{t,k+1} | \mathcal{F}_{t,k}\right] = \frac{1}{n} \sum_{i=1}^n \mu_{t,k+1,i} \, ,$$

$$\mathbb{E}\left[\| \eta_{t,k+1} - \mathbb{E}\left[\eta_{t,k+1} | \mathcal{F}_{t,k}\right] \|^2 | \mathcal{F}_{t,k}\right] \leq \frac{1}{\mathsf{b}} \left( \frac{C_v}{M_{t,k+1}} + \frac{C_{vb}^2}{\bar{M}_{t,k+1}^2} \right) \, .$$

*Proof* Let $t \in [k^{\text{out}}]^\star$ and $k \in [k_t^{\text{in}} - 1]$. We have

$$\mathbb{E}\left[\eta_{t,k+1} | \mathcal{F}_{t,k+1/2}\right] = \frac{1}{\mathsf{b}} \sum_{i \in \mathcal{B}_{t,k+1}} \mu_{t,k+1,i} \, ,$$

since $\mathcal{B}_{t,k+1} \in \mathcal{F}_{t,k+1/2}$; and by Lemma 7.1,

$$\mathbb{E}\left[\eta_{t,k+1} | \mathcal{F}_{t,k}\right] = \frac{1}{n} \sum_{i=1}^n \mu_{t,k+1,i} \, .$$

We write

$$\begin{aligned} &\eta_{t,k+1} - \mathbb{E}\left[\eta_{t,k+1} | \mathcal{F}_{t,k}\right] \\ &\quad = \frac{1}{\mathsf{b}} \sum_{i \in \mathcal{B}_{t,k+1}} \xi_{t,k+1,i} - \frac{1}{n} \sum_{i=1}^n \mu_{t,k+1,i} \\ &\quad = \frac{1}{\mathsf{b}} \sum_{i \in \mathcal{B}_{t,k+1}} \left\{\xi_{t,k+1,i} - \mu_{t,k+1,i}\right\} \\ &\quad + \frac{1}{\mathsf{b}} \sum_{i \in \mathcal{B}_{t,k+1}} \mu_{t,k+1,i} - \frac{1}{n} \sum_{i=1}^n \mu_{t,k+1,i} \, . \end{aligned}$$

The RHS is of the form $U + V$ and we write $\|U + V\|^2 = \|U\|^2 + \|V\|^2 + 2\langle U, V \rangle$ with $U \leftarrow \mathsf{b}^{-1} \sum_{i \in \mathcal{B}_{t,k+1}} \left\{\xi_{t,k+1,i} - \mu_{t,k+1,i}\right\}$. By conditioning and by definition of $\sigma_{t,k+1,i}^2$, we have

$$\mathbb{E}\left[\|U\|^2 | \mathcal{F}_{t,k}\right] = \frac{1}{\mathsf{b}^2} \mathbb{E}\left[\sum_{i \in \mathcal{B}_{t,k+1}} \sigma_{t,k+1,i}^2 | \mathcal{F}_{t,k}\right] \, .$$

Under A4-item 1, we have by Lemma 7.1

$$\mathbb{E}\left[\|U\|^2 | \mathcal{F}_{t,k}\right] = \frac{1}{\mathsf{b}\, n} \sum_{i=1}^n \sigma_{t,k+1,i}^2 \leq \frac{C_v}{\mathsf{b}\, M_{t,k+1}} \, .$$

By Lemma 7.1 again, it holds

$$\mathbb{E}\left[\|V\|^2 | \mathcal{F}_{t,k}\right] \leq \frac{1}{\mathsf{b}\, n} \sum_{i=1}^n \| \mu_{t,k+1,i} - \frac{1}{n} \sum_{j=1}^n \mu_{t,k+1,j} \|^2 \, ,$$

which yields

$$\mathbb{E}\left[\|V\|^2 | \mathcal{F}_{t,k}\right] \leq \frac{C_{vb}^2}{\mathsf{b}\, \bar{M}_{t,k+1}^2} \, .$$

Finally, upon noting that $\mathbb{E}\left[U | \mathcal{F}_{t,k+1/2}\right] = 0$ and $V \in \mathcal{F}_{t,k+1/2}$, we have

$$\mathbb{E}\left[\langle U, V \rangle | \mathcal{F}_{t,k}\right] = \mathbb{E}\left[\left\langle \mathbb{E}\left[U | \mathcal{F}_{t,k+1/2}\right], V \right\rangle | \mathcal{F}_{t,k}\right] = 0 \, .$$

This concludes the proof. □

## 7.3 Results on the variables $\mathsf{S}_{t,k}$

Proposition 7.3 studies the bias of the variables $\mathsf{S}_{t,k+1}$. It shows that $\mathsf{S}_{t,k+1}$ is a *biased* approximation of $\bar{\mathsf{h}}(\widehat{S}_{t,k})$:

$$\mathbb{E}\left[\mathsf{S}_{t,k+1} | \mathcal{F}_{t,k}\right] \neq \bar{\mathsf{h}}(\widehat{S}_{t,k}).$$

When $k=0$, we may have $\mathbb{E}\left[\mathsf{S}_{t,1}|\mathcal{F}_{t,0}\right]=\bar{\mathsf{h}}(\widehat{S}_{t,0})$ if $\delta_{t,0,i}=\mathsf{h}_i(\widehat{S}_{t,0})$ and $\mathcal{B}_{t,0}=[n]^\star$. The choice $\mathcal{B}_{t,0}=[n]^\star$ is the strategy proposed in Wang et al (2019) for `SpiderBoost`; it has an important computational cost but has the advantage to cancel the bias of the variable $\mathsf{S}_\cdot$ at the beginning of each outer loop. Along the inner loops, a (signed) bias appears.

**Proposition 7.3** *For any $t\in[k^{\rm out}]^\star$ and $k\in[k_t^{\rm in}-1]$, it holds*

$$\mathbb{E}\left[\mathsf{S}_{t,k+1}|\mathcal{F}_{t,k}\right]-\bar{\mathsf{h}}(\widehat{S}_{t,k}) = \mathsf{S}_{t,k}-\bar{\mathsf{h}}(\widehat{S}_{t,k-1})+\mathbb{E}\left[\eta_{t,k+1}|\mathcal{F}_{t,k}\right]\ ,$$

*and*

$$\mathbb{E}\left[\mathsf{S}_{t,k+1}-\bar{\mathsf{h}}(\widehat{S}_{t,k})|\mathcal{F}_{t,0}\right]=\mathcal{E}_t+\sum_{j=1}^{k+1}\mathbb{E}\left[\eta_{t,j}|\mathcal{F}_{t,0}\right]\ .$$

*Proof* Let $t\in[k^{\rm out}]^\star$ and $k\in[k_t^{\rm in}-1]$. We write $\mathsf{S}_{t,k+1}=\mathsf{S}_{t,k}+h_{\mathcal{B}_{t,k+1}}(\widehat{S}_{t,k})-h_{\mathcal{B}_{t,k+1}}(\widehat{S}_{t,k-1})+\eta_{t,k+1}$. By Lemma 7.1,

$$\mathbb{E}\left[\mathsf{S}_{t,k+1}|\mathcal{F}_{t,k}\right]=\mathsf{S}_{t,k}+\bar{\mathsf{h}}(\widehat{S}_{t,k})-\bar{\mathsf{h}}(\widehat{S}_{t,k-1}) +\mathbb{E}\left[\eta_{t,k+1}|\mathcal{F}_{t,k}\right]\ .$$

Since $\mathcal{F}_{t,0}\subseteq\mathcal{F}_{t,k}$, we have

$$\mathbb{E}\left[\mathsf{S}_{t,k+1}-\bar{\mathsf{h}}(\widehat{S}_{t,k})|\mathcal{F}_{t,0}\right]=\mathbb{E}\left[\mathsf{S}_{t,k}-\bar{\mathsf{h}}(\widehat{S}_{t,k-1})|\mathcal{F}_{t,0}\right] +\mathbb{E}\left[\eta_{t,k+1}|\mathcal{F}_{t,0}\right]\ .$$

Summing from $j=0$ to $j=k$ yields

$$\mathbb{E}\left[\mathsf{S}_{t,k+1}-\bar{\mathsf{h}}(\widehat{S}_{t,k})|\mathcal{F}_{t,0}\right]=\mathsf{S}_{t,0}-\bar{\mathsf{h}}(\widehat{S}_{t,-1}) +\sum_{j=0}^{k}\mathbb{E}\left[\eta_{t,j+1}|\mathcal{F}_{t,0}\right]\ .$$

The proof is concluded by using $\widehat{S}_{t,0}=\widehat{S}_{t,-1}$ and the definition of $\mathcal{E}_t$; note that $\mathcal{E}_t\in\mathcal{F}_{t,0}$. □

Proposition 7.4 provides a control of the conditional variance of $\mathsf{S}_{t,k}$.

**Proposition 7.4** *Assume A 2, A 4-item 1 and A 4-item 3. For any $t\in[k^{\rm out}]^\star$ and $k\in[k_t^{\rm in}-1]$, it holds*

$$\mathbb{E}\left[\left\|\mathsf{S}_{t,k+1}-\mathbb{E}\left[\mathsf{S}_{t,k+1}|\mathcal{F}_{t,k}\right]\right\|^2|\mathcal{F}_{t,k}\right] \le\frac{L^2}{\mathsf{b}}\left(1+\frac{2C_{vb}}{\sqrt{\mathsf{b}}\,\bar{M}_{t,k+1}}\right)\|\widehat{S}_{t,k}-\widehat{S}_{t,k-1}\|^2 +\frac{C_v}{\mathsf{b}\,M_{t,k+1}}+\frac{C_{vb}^2}{\mathsf{b}\,\bar{M}_{t,k+1}^2}+\frac{2C_{vb}}{\sqrt{\mathsf{b}}\,\bar{M}_{t,k+1}}\ .$$

*Proof* Let $t\in[k^{\rm out}]^\star$, $k\in[k_t^{\rm in}-1]$. By Lemma 7.1, Proposition 7.3, the definitions of $\mathsf{S}_{t,k+1}$ and of the filtration $\mathcal{F}_{t,k}$,

$$\begin{aligned}&\mathsf{S}_{t,k+1}-\mathbb{E}\left[\mathsf{S}_{t,k+1}|\mathcal{F}_{t,k}\right]\\ &=\mathsf{S}_{t,k+1}-\bar{\mathsf{h}}(\widehat{S}_{t,k})-\mathsf{S}_{t,k}+\bar{\mathsf{h}}(\widehat{S}_{t,k-1})-\mathbb{E}\left[\eta_{t,k+1}|\mathcal{F}_{t,k}\right]\\ &=\eta_{t,k+1}-\mathbb{E}\left[\eta_{t,k+1}|\mathcal{F}_{t,k}\right]\\ &+\bar{\mathsf{h}}_{\mathcal{B}_{t,k+1}}(\widehat{S}_{t,k})-\bar{\mathsf{h}}_{\mathcal{B}_{t,k+1}}(\widehat{S}_{t,k-1})-\bar{\mathsf{h}}(\widehat{S}_{t,k})+\bar{\mathsf{h}}(\widehat{S}_{t,k-1})\ .\end{aligned}$$

The RHS is of the form $U+V$ with $U\leftarrow\eta_{t,k+1}-\mathbb{E}\left[\eta_{t,k+1}|\mathcal{F}_{t,k}\right]$ and $V\in\mathcal{F}_{t,k+1/2}$. Then, we write $\mathbb{E}\left[\|U+V\|^2|\mathcal{F}_{t,k}\right]=\mathbb{E}\left[\|U\|^2|\mathcal{F}_{t,k}\right]+\mathbb{E}\left[\|V\|^2|\mathcal{F}_{t,k}\right]+2\mathbb{E}\left[\left\langle\mathbb{E}\left[U|\mathcal{F}_{t,k+1/2}\right],V\right\rangle|\mathcal{F}_{t,k}\right]$.

The term $\mathbb{E}\left[\|V\|^2|\mathcal{F}_{t,k}\right]$ is controlled by Lemma 7.1: an upper bound is $L^2\mathsf{b}^{-1}\|\widehat{S}_{t,k}-\widehat{S}_{t,k-1}\|^2$. The term $\mathbb{E}\left[\|U\|^2|\mathcal{F}_{t,k}\right]$ is controlled by Lemma 7.2: an upper bound is $C_v/(\mathsf{b}\,M_{t,k+1})+C_{vb}^2/(\mathsf{b}\,\bar{M}_{t,k+1}^2)$. Upon noting that $V\in\mathcal{F}_{t,k+1/2}$, and using Lemma 7.2 and Lemma 7.1, the scalar product is upper bounded by

$$\begin{aligned}&2\,\mathbb{E}\left[\|V\|\left\|\mathbb{E}\left[U|\mathcal{F}_{t,k+1/2}\right]\right\|\middle|\mathcal{F}_{t,k}\right]\\ &\le 2\frac{C_{vb}}{\sqrt{\mathsf{b}}\,\bar{M}_{t,k+1}}\left\{\mathbb{E}\left[\|V\|^2\middle|\mathcal{F}_{t,k}\right]\right\}^{1/2}\\ &\le 2\frac{C_{vb}}{\sqrt{\mathsf{b}}\,\bar{M}_{t,k+1}}\left(1+\mathbb{E}\left[\|V\|^2|\mathcal{F}_{t,k}\right]\right)\ ,\end{aligned}$$

where we used that $\mathsf{a}\le1+\mathsf{a}^2$. □

Proposition 7.5 establishes an upper bound on the conditional expectation of the quadratic error $\|\mathsf{S}_{t,k+1}-\bar{\mathsf{h}}(\widehat{S}_{t,k})\|^2$.

**Proposition 7.5** *Assume A 2 and A 4. For any $t\in[k^{\rm out}]^\star$ and $k\in[k_t^{\rm in}-1]$, it holds*

$$\begin{aligned}&\mathbb{E}\left[\|\mathsf{S}_{t,k+1}-\bar{\mathsf{h}}(\widehat{S}_{t,k})\|^2|\mathcal{F}_{t,k}\right]\\ &\le\left(1+\frac{2C_b}{m_{t,k+1}}\right)\|\mathsf{S}_{t,k}-\bar{\mathsf{h}}(\widehat{S}_{t,k-1})\|^2\\ &+\frac{L^2}{\mathsf{b}}\left(1+\frac{2C_{vb}}{\sqrt{\mathsf{b}}\,\bar{M}_{t,k+1}}\right)\|\widehat{S}_{t,k}-\widehat{S}_{t,k-1}\|^2\\ &+\mathcal{U}_{t,k+1}\ ,\end{aligned}$$

*where*

$$\mathcal{U}_{t,k}\stackrel{\rm def}{=}\frac{2C_b}{m_{t,k}}+\frac{C_b^2}{m_{t,k}^2}+\frac{C_v}{\mathsf{b}\,M_{t,k}}+\frac{2C_{vb}}{\sqrt{\mathsf{b}}\,\bar{M}_{t,k}}+\frac{C_{vb}^2}{\mathsf{b}\,\bar{M}_{t,k}^2}\ .$$

*Proof* Let $t \in [k^{\rm out}]^\star$ and $k \in [k_t^{\rm in}-1]$. By definition of the conditional expectation, we have for any r.v. $U, V$ and any $\sigma$-field $\mathcal{F}$ such that $V \in \mathcal{F}$:

$$\mathbb{E}\left[\|U-V\|^2|\mathcal{F}\right] = \mathbb{E}\left[\|U-\mathbb{E}\left[U|\mathcal{F}\right]\|^2|\mathcal{F}\right] + \|\mathbb{E}\left[U|\mathcal{F}\right]-V\|^2 .$$

We apply this equality with $U \leftarrow \widehat{S}_{t,k+1}$, $V \leftarrow \bar{\mathsf{h}}(\widehat{S}_{t,k})$ and $\mathcal{F} \leftarrow \mathcal{F}_{t,k}$. Proposition 7.4 controls the first term. For the second one, by Proposition 7.3, Lemma 7.2 and A4-item 2 we have

$$\begin{aligned}
&\|\mathbb{E}\left[\mathsf{S}_{t,k+1}|\mathcal{F}_{t,k}\right] - \bar{\mathsf{h}}(\widehat{S}_{t,k})\|^2 \\
&= \|\mathsf{S}_{t,k} - \bar{\mathsf{h}}(\widehat{S}_{t,k-1}) + \mathbb{E}\left[\eta_{t,k+1}|\mathcal{F}_{t,k}\right]\|^2 \\
&\le \|\mathsf{S}_{t,k} - \bar{\mathsf{h}}(\widehat{S}_{t,k-1})\|^2 + \frac{C_b^2}{m_{t,k+1}^2} \\
&+ 2\frac{C_b}{m_{t,k+1}}\|\mathsf{S}_{t,k} - \bar{\mathsf{h}}(\widehat{S}_{t,k-1})\| .
\end{aligned}$$

We conclude by using $\|a\| \le 1+\|a\|^2$ with $a \leftarrow \|\mathsf{S}_{t,k} - \bar{\mathsf{h}}(\widehat{S}_{t,k-1})\|$. □

**Corollary 7.6** (of Proposition 7.5) *Assume also A3-item 3. For $t \in [k^{\rm out}]^\star$ and $k \in [k_t^{\rm in}-1]$, define $\mathcal{D}_{t,k+1}$ by*

$$\|\widehat{S}_{t,k+1} - \mathrm{prox}^{B_{t,k}}_{\gamma_{t,k+1}g}(\widehat{S}_{t,k} + \gamma_{t,k+1}\bar{\mathsf{h}}(\widehat{S}_{t,k}))\|^2_{B_{t,k}} ,$$

*and $\mathcal{D}_{t,0} \overset{\rm def}{=} 0$. For $t \in [k^{\rm out}]^\star$ and $k \in [k_t^{\rm in}-]$, it holds*

$$\begin{aligned}
&\mathbb{E}\left[\|\mathsf{S}_{t,k+1} - \bar{\mathsf{h}}(\widehat{S}_{t,k})\|^2|\mathcal{F}_{t,k}\right] \\
&\le \left(1+\frac{2C_b}{m_{t,k+1}}\right)\|\mathsf{S}_{t,k} - \bar{\mathsf{h}}(\widehat{S}_{t,k-1})\|^2 \\
&+ \gamma_{t,k}^2\frac{2}{v_{\min}}\frac{L^2}{\mathsf{b}}\left(1+\frac{2C_{vb}}{\sqrt{\mathsf{b}}\,\bar{M}_{t,k+1}}\right)\Delta^\star_{t,k} \\
&+ \frac{2}{v_{\min}}\frac{L^2}{\mathsf{b}}\left(1+\frac{2C_{vb}}{\sqrt{\mathsf{b}}\,\bar{M}_{t,k+1}}\right)\mathcal{D}_{t,k} \\
&+ \mathcal{U}_{t,k+1} .
\end{aligned}$$

*By convention, $\Delta^\star_{t,0} \overset{\rm def}{=} 0$.*

*Proof* The proof consists in an upper bound for $\|\widehat{S}_{t,k} - \widehat{S}_{t,k-1}\|^2$. Let $s \in \mathcal{S}$, $H, h \in \mathbb{R}^q$, $\gamma > 0$ and $B$ be a $q \times q$ positive definite matrix. For any $\beta > 0$, it holds

$$\begin{aligned}
&\|\mathrm{prox}^B_{\gamma g}(s+\gamma H) - s\|^2_B \le (1+\frac{1}{\beta})\,\|\mathrm{prox}^B_{\gamma g}(s+\gamma h) - s\|^2_B \\
&+ (1+\beta)\,\|\mathrm{prox}^B_{\gamma g}(s+\gamma H) - \mathrm{prox}^B_{\gamma g}(s+\gamma h)\|^2_B .
\end{aligned}$$

We apply these inequalities with $\gamma \leftarrow \gamma_{t,k}$, $B \leftarrow B_{t,k-1}$, $s \leftarrow \widehat{S}_{t,k-1}$, $H \leftarrow \mathsf{S}_{t,k}$ and $h \leftarrow \bar{\mathsf{h}}(\widehat{S}_{t,k-1})$. Then, for any $k > 0$,

$$\begin{aligned}
&\|\widehat{S}_{t,k} - \widehat{S}_{t,k-1}\|^2_{B_{t,k-1}} \\
&\le (1+\beta^{-1})\,\gamma^2_{t,k}\Delta^\star_{t,k} + (1+\beta)\mathcal{D}_{t,k} . \qquad (24)
\end{aligned}$$

We choose $\beta = 1$ and conclude by A3-item 3: $\|\cdot\|^2 \le v_{\min}^{-1}\|\cdot\|^2_{B_{t,k-1}}$.

When $k = 0$, $\|\widehat{S}_{t,k} - \widehat{S}_{t,k-1}\|^2_{B_{t,k-1}} = 0$ since by definition, $\widehat{S}_{t,0} = \widehat{S}_{t,-1}$. Therefore, (24) remains valid since $\mathcal{D}_{t,0} = 0$ and $\Delta^\star_{t,0} = 0$ by convention. This concludes the proof. □

**Corollary 7.7** (of Corollary 7.6) *Let $\{\rho_{t,k}, t \ge 1, k \ge 0\}$ be a positive sequence satisfying*

$$\rho_{t,k+1}\left(1+\frac{2C_b}{m_{t,k+1}}\right) \le \rho_{t,k} . \qquad (25)$$

*For any $t \in [k^{\rm out}]^\star$, $k \in [k_t^{\rm in}-1]$, it holds*

$$\begin{aligned}
&\rho_{t,k+1}\mathbb{E}\left[\|\mathsf{S}_{t,k+1} - \bar{\mathsf{h}}(\widehat{S}_{t,k})\|^2|\mathcal{F}_{t,0}\right] \\
&\le \rho_{t,0}\|\mathcal{E}_t\|^2 + \sum_{\ell=1}^{k+1}\rho_{t,\ell}\,\mathcal{U}_{t,\ell} + \frac{2}{v_{\min}}\frac{L^2}{\mathsf{b}}\cdots \\
&\times\left\{\sum_{\ell=1}^{k}\gamma^2_{t,\ell}\rho_{t,\ell+1}\left(1+\frac{2C_{vb}}{\sqrt{\mathsf{b}}\,\bar{M}_{t,\ell+1}}\right)\mathbb{E}\left[\Delta^\star_{t,\ell}|\mathcal{F}_{t,0}\right]\right. \\
&\left.+\sum_{\ell=1}^{k}\rho_{t,\ell+1}\left(1+\frac{2C_{vb}}{\sqrt{\mathsf{b}}\,\bar{M}_{t,\ell+1}}\right)\mathbb{E}\left[\mathcal{D}_{t,\ell}|\mathcal{F}_{t,0}\right]\right\} .
\end{aligned}$$

*Proof* In Corollary 7.6, the claim is of the form

$$\begin{aligned}
&\mathbb{E}\left[\|\mathsf{S}_{t,k+1} - \bar{\mathsf{h}}(\widehat{S}_{t,k})\|^2|\mathcal{F}_{t,k}\right] \\
&\le \left(1+\frac{2C_b}{m_{t,k+1}}\right)\|\mathsf{S}_{t,k} - \bar{\mathsf{h}}(\widehat{S}_{t,k-1})\|^2 + A_k .
\end{aligned}$$

This yields, by using the condition (25),

$$\begin{aligned}
&\rho_{t,k+1}\mathbb{E}\left[\|\mathsf{S}_{t,k+1} - \bar{\mathsf{h}}(\widehat{S}_{t,k})\|^2|\mathcal{F}_{t,k}\right] \\
&\le \rho_{t,k}\,\|\mathsf{S}_{t,k} - \bar{\mathsf{h}}(\widehat{S}_{t,k-1})\|^2 + \rho_{t,k+1}\,A_k .
\end{aligned}$$

Using $\mathbb{E}[U|\mathcal{F}_{t,0}] = \mathbb{E}\left[\mathbb{E}[U|\mathcal{F}_{t,k}]|\mathcal{F}_{t,0}\right]$ and summing from $\ell = 0$ to $\ell = k$ yields

$$\begin{aligned}
&\rho_{t,k+1}\,\mathbb{E}\left[\|\mathsf{S}_{t,k+1} - \bar{\mathsf{h}}(\widehat{S}_{t,k})\|^2|\mathcal{F}_{t,0}\right] \\
&\le \sum_{\ell=0}^{k}\rho_{t,\ell+1}\mathbb{E}\left[A_\ell|\mathcal{F}_{t,0}\right] + \rho_{t,0}\|\mathsf{S}_{t,0} - \bar{\mathsf{h}}(\widehat{S}_{t,-1})\|^2 ;
\end{aligned}$$

we then conclude by using the equality $\widehat{S}_{t,-1} = \widehat{S}_{t,0}$ and the definition of $\mathcal{E}_t$. Note also that $\Delta^\star_{t,0} = 0$ and $\mathcal{D}_{t,0} = 0$. □

## 7.4 Lyapunov inequalities for W, $g$ and W $+g$

Lemma 7.8, while being classical in smooth optimization, is provided for a self-content purpose.

**Lemma 7.8** *Assume A3. For any $s, s' \in \mathcal{S}$ and $\gamma > 0$,*

$$\mathrm{W}(s') \leq \mathrm{W}(s) - \left\langle \bar{\mathsf{h}}(s), s' - s \right\rangle_{B(s)} + \frac{L_{\dot{\mathrm{W}}}}{2} \|s' - s\|^2 \ .$$

*Proof* $\mathcal{S}$ is convex since it is the domain of a convex function. By A3, W is continuously differentiable on $\mathcal{S}$ with $L_{\dot{\mathrm{W}}}$-Lipschitz gradient. Then for any $s, s' \in \mathcal{S}$,

$$\mathrm{W}(s') - \mathrm{W}(s) \leq \left\langle \nabla \mathrm{W}(s), s' - s \right\rangle + \frac{L_{\dot{\mathrm{W}}}}{2} \|s' - s\|^2 \ .$$

We use that $\nabla \mathrm{W}(s) = -B(s)\bar{\mathsf{h}}(s)$, so that

$$\left\langle \nabla \mathrm{W}(s), s' - s \right\rangle = -\left\langle \bar{\mathsf{h}}(s), s' - s \right\rangle_{B(s)} \ .$$

□

**Lemma 7.9** *Assume A1. Let $B$ be a $q \times q$ positive definite matrix. For any $s \in \mathcal{S}$, $\gamma > 0$, $H, h \in \mathbb{R}^q$ and $\beta > 0$,*

$$\begin{aligned} g\left(\mathrm{prox}^B_{\gamma g}(s + \gamma H)\right) &\leq g(s) \\ &- \frac{1}{\gamma}\left(1 - \frac{\beta}{4}\right) \|\mathrm{prox}^B_{\gamma g}(s + \gamma h) - s\|^2_B \\ &- \frac{1}{\gamma}(1 - \frac{1}{\beta})\|\mathrm{prox}^B_{\gamma g}(s + \gamma h) - \mathrm{prox}^B_{\gamma g}(s + \gamma H)\|^2_B \\ &- \left\langle h, s - \mathrm{prox}^B_{\gamma g}(s + \gamma h) \right\rangle_B \\ &+ \left\langle H, \mathrm{prox}^B_{\gamma g}(s + \gamma H) - \mathrm{prox}^B_{\gamma g}(s + \gamma h) \right\rangle_B \ . \end{aligned}$$

*Proof* In this proof, we use the shorthand notation $\mathsf{p}_H \overset{\text{def}}{=} \mathrm{prox}^B_{\gamma g}(s + \gamma H)$ and $\mathsf{p}_h \overset{\text{def}}{=} \mathrm{prox}^B_{\gamma g}(s + \gamma h)$. By Lemma 3.1 and the definition of the subdifferential at a point, it holds

$$\begin{aligned} g(\mathsf{p}_h) &\geq g(\mathsf{p}_H) - \gamma^{-1} \left\langle \mathsf{p}_H - s - \gamma H, \mathsf{p}_h - \mathsf{p}_H \right\rangle_B \\ g(s) &\geq g(\mathsf{p}_h) - \gamma^{-1} \left\langle \mathsf{p}_h - s - \gamma h, s - \mathsf{p}_h \right\rangle_B \ . \end{aligned}$$

This yields

$$\begin{aligned} g(\mathsf{p}_H) &\leq g(s) - \gamma^{-1}\|\mathsf{p}_h - s\|^2_B - \langle h, s - \mathsf{p}_h \rangle_B \\ &\quad - \langle H, \mathsf{p}_h - \mathsf{p}_H \rangle_B + \gamma^{-1} \left\langle \mathsf{p}_H - s, \mathsf{p}_h - \mathsf{p}_H \right\rangle_B \ . \end{aligned}$$

For the last term, we write for any $\beta > 0$,

$$\begin{aligned} &\gamma^{-1} \left\langle \mathsf{p}_H - s, \mathsf{p}_h - \mathsf{p}_H \right\rangle_B + \gamma^{-1}\|\mathsf{p}_h - \mathsf{p}_H\|^2_B \\ &= \gamma^{-1} \left\langle \mathsf{p}_h - s, \mathsf{p}_h - \mathsf{p}_H \right\rangle_B \\ &\leq 2 \left\langle (\mathsf{p}_h - s)\frac{\sqrt{\beta}}{2\sqrt{\gamma}}, (\mathsf{p}_h - \mathsf{p}_H)\frac{1}{\sqrt{\beta\gamma}} \right\rangle_B \\ &\leq \frac{\beta}{4\gamma}\|\mathsf{p}_h - s\|^2_B + \frac{1}{\beta\gamma}\|\mathsf{p}_h - \mathsf{p}_H\|^2_B \ . \end{aligned}$$

This concludes the proof. □

**Proposition 7.10** *Assume A1, A2 and A3. For any $t \in [k^{\mathrm{out}}]^\star$, $k \in [k_t^{\mathrm{in}} - 1]$ and $\beta > 0$,*

$$\begin{aligned} &\mathbb{E}\left[\mathrm{W}(\widehat{S}_{t,k+1}) + g(\widehat{S}_{t,k+1})|\mathcal{F}_{t,0}\right] \\ &\leq \mathbb{E}\left[\mathrm{W}(\widehat{S}_{t,k}) + g(\widehat{S}_{t,k})|\mathcal{F}_{t,0}\right] \\ &- \gamma_{t,k+1}\left(1 - \frac{\beta}{4} - \frac{L_{\dot{\mathrm{W}}}\gamma_{t,k+1}}{v_{\min}}\right)\mathbb{E}\left[\Delta^\star_{t,k+1}|\mathcal{F}_{t,0}\right] \\ &- \frac{1}{\gamma_{t,k+1}}\left(1 - \frac{1}{\beta} - \frac{L_{\dot{\mathrm{W}}}}{v_{\min}}\gamma_{t,k+1}\right)\mathbb{E}\left[\mathcal{D}_{t,k+1}|\mathcal{F}_{t,0}\right] \\ &+ \gamma_{t,k+1}\mathbb{E}\left[\|\mathsf{S}_{t,k+1} - \bar{\mathsf{h}}(\widehat{S}_{t,k})\|^2_{B_{t,k}}|\mathcal{F}_{t,0}\right] \ , \end{aligned}$$

*where $\mathcal{D}_{t,k+1}$ is defined in Corollary 7.6.*

*Proof* Let $\gamma > 0$, $s \in \mathcal{S}$ and $H \in \mathbb{R}^q$. Apply Lemma 7.8 with $s' \leftarrow \mathrm{prox}^{B(s)}_{\gamma g}(s + \gamma H) \in \mathcal{S}$; and Lemma 7.9 with $h \leftarrow \bar{\mathsf{h}}(s)$. This yields for any $\beta > 0$,

$$\begin{aligned} &\mathrm{W}(\mathrm{prox}^B_{\gamma g}(s + \gamma H)) + g(\mathrm{prox}^B_{\gamma g}(s + \gamma H)) \\ &\leq \mathrm{W}(s) + g(s) - \frac{1}{\gamma}(1 - \frac{\beta}{4})\|\mathrm{prox}^B_{\gamma g}(s + \gamma\bar{\mathsf{h}}(s)) - s\|^2_B \\ &- \frac{1}{\gamma}\left(1 - \frac{1}{\beta}\right)\|\mathrm{prox}^B_{\gamma g}(s + \gamma H) - \mathrm{prox}^B_{\gamma g}(s + \gamma\bar{\mathsf{h}}(s))\|^2_B \\ &- \left\langle \bar{\mathsf{h}}(s) - H, \mathrm{prox}^B_{\gamma g}(s + \gamma H) - \mathrm{prox}^B_{\gamma g}(s + \gamma\bar{\mathsf{h}}(s)) \right\rangle_B \\ &\qquad + \frac{L_{\dot{\mathrm{W}}}}{2}\|\mathrm{prox}^B_{\gamma g}(s + \gamma H) - s\|^2 \ . \end{aligned}$$

Since $\mathrm{prox}^B_{\gamma g}$ is firmly nonexpansive (see Lemma 6.1), the scalar product is upper bounded by $\gamma\|H - \bar{\mathsf{h}}(s)\|^2_B$. By A3-item 3, we write

$$\|\mathrm{prox}^B_{\gamma g}(s+\gamma H) - s\|^2 \leq \frac{1}{v_{\min}}\|\mathrm{prox}^B_{\gamma g}(s+\gamma H) - s\|^2_B \ ;$$

then we use $\|a + b\|^2_B \leq 2\|a\|^2_B + 2\|b\|^2_B$ with $a \leftarrow \mathrm{prox}^B_{\gamma g}(s + \gamma H) - \mathrm{prox}^B_{\gamma g}(s + \gamma\bar{\mathsf{h}}(s))$. This yields

$$\begin{aligned} &\frac{L_{\dot{\mathrm{W}}}}{2}\|\mathrm{prox}^B_{\gamma g}(s + \gamma H) - s\|^2 \\ &\quad \leq \frac{L_{\dot{\mathrm{W}}}}{v_{\min}}\|\mathrm{prox}^B_{\gamma g}(s + \gamma H) - \mathrm{prox}^B_{\gamma g}(s + \gamma\bar{\mathsf{h}}(s))\|^2_B \\ &\quad + \frac{L_{\dot{\mathrm{W}}}}{v_{\min}}\|\mathrm{prox}^B_{\gamma g}(s + \gamma\bar{\mathsf{h}}(s)) - s\|^2_B \ . \end{aligned}$$

We apply these inequalities with $s \leftarrow \widehat{S}_{t,k}$, $\gamma \leftarrow \gamma_{t,k+1}$, $H \leftarrow \mathsf{S}_{t,k+1}$, $s' \leftarrow \widehat{S}_{t,k+1}$ and $B \leftarrow B_{t,k}$. Note that $\mathrm{prox}^B_{\gamma g}(s+\gamma H) = \widehat{S}_{t,k+1}$. The proof is concluded. □

## 7.5 Proof of Theorem 4.1

Let $t \in [k^{\mathrm{out}}]^\star$. Let $\mu \in (0,1)$. Throughout the proof, set

$$A_{t,k+1} \overset{\mathrm{def}}{=} \left(1 + \frac{2C_{vb}}{\sqrt{\mathsf{b}}\,\bar{M}_{t,k+1}}\right) .$$

From Corollary 7.7 applied with $\rho_{t,k+1} \leftarrow \gamma_{t,k+1}$ and Proposition 7.10 applied with $\beta \leftarrow 4\mu$, it holds for any $k \in [k_t^{\mathrm{in}} - 1]$,

$$\begin{aligned}
&\mathbb{E}\left[\mathrm{W}(\widehat{S}_{t,k+1}) + g(\widehat{S}_{t,k+1})|\mathcal{F}_{t,0}\right] \\
&\le \mathbb{E}\left[\mathrm{W}(\widehat{S}_{t,k}) + g(\widehat{S}_{t,k})|\mathcal{F}_{t,0}\right] \\
&- \gamma_{t,k+1}\left(1 - \mu - \frac{L_{\dot{\mathrm{W}}}}{v_{\min}}\gamma_{t,k+1}\right)\mathbb{E}\left[\Delta^\star_{t,k+1}|\mathcal{F}_{t,0}\right] \\
&- \frac{1}{\gamma_{t,k+1}}\left(1 - \frac{1}{4\mu} - \frac{L_{\dot{\mathrm{W}}}}{v_{\min}}\gamma_{t,k+1}\right)\mathbb{E}\left[\mathcal{D}_{t,k+1}|\mathcal{F}_{t,0}\right] \\
&+ \gamma_{t,0} v_{\max}\|\mathcal{E}_t\|^2 + v_{\max}\sum_{\ell=1}^{k+1}\gamma_{t,\ell}\,\mathcal{U}_{t,\ell} \\
&+ \frac{2v_{\max}}{v_{\min}}\frac{L^2}{\mathsf{b}}\sum_{\ell=1}^{k}\gamma_{t,\ell}^3 A_{t,\ell+1}\mathbb{E}\left[\Delta^\star_{t,\ell}|\mathcal{F}_{t,0}\right] \\
&+ \frac{2v_{\max}}{v_{\min}}\frac{L^2}{\mathsf{b}}\sum_{\ell=1}^{k}\gamma_{t,\ell+1} A_{t,\ell+1}\mathbb{E}\left[\mathcal{D}_{t,\ell}|\mathcal{F}_{t,0}\right] .
\end{aligned}$$

Above, we used that $\gamma_{t,k+1} \le \gamma_{t,\ell}$ for any $\ell \in [k+1]$. We now sum from $k = 0$ to $k = k_t^{\mathrm{in}} - 1$. This yields,

$$\begin{aligned}
&\mathbb{E}\left[\mathrm{W}(\widehat{S}_{t,k_t^{\mathrm{in}}}) + g(\widehat{S}_{t,k_t^{\mathrm{in}}})|\mathcal{F}_{t,0}\right] \\
&\quad\le \mathbb{E}\left[\mathrm{W}(\widehat{S}_{t,0}) + g(\widehat{S}_{t,0})|\mathcal{F}_{t,0}\right] \\
&- \sum_{k=1}^{k_t^{\mathrm{in}}}\gamma_{t,k}\left(1 - \mu - \frac{L_{\dot{\mathrm{W}}}}{v_{\min}}\gamma_{t,k}\right)\mathbb{E}\left[\Delta^\star_{t,k}|\mathcal{F}_{t,0}\right] \\
&- \sum_{k=1}^{k_t^{\mathrm{in}}}\frac{1}{\gamma_{t,k}}\left(1 - \frac{1}{4\mu} - \frac{L_{\dot{\mathrm{W}}}}{v_{\min}}\gamma_{t,k}\right)\mathbb{E}\left[\mathcal{D}_{t,k}|\mathcal{F}_{t,0}\right] \\
&+ \gamma_{t,0} v_{\max} k_t^{\mathrm{in}}\|\mathcal{E}_t\|^2 \\
&+ v_{\max}\sum_{\ell=1}^{k_t^{\mathrm{in}}}(k_t^{\mathrm{in}} - \ell + 1)\gamma_{t,\ell}\,\mathcal{U}_{t,\ell} \\
&+ \frac{2v_{\max}}{v_{\min}} k_t^{\mathrm{in}}\sum_{k=1}^{k_t^{\mathrm{in}}-1}\gamma_{t,k}^3 A_{t,k+1}\mathbb{E}\left[\Delta^\star_{t,k}|\mathcal{F}_{t,0}\right] \\
&+ \frac{2v_{\max}}{v_{\min}} k_t^{\mathrm{in}}\sum_{k=1}^{k_t^{\mathrm{in}}-1}\gamma_{t,k+1} A_{t,k+1}\mathbb{E}\left[\mathcal{D}_{t,k}|\mathcal{F}_{t,0}\right] .
\end{aligned}$$

Observe that the coefficient in front of $\mathbb{E}\left[\Delta^\star_{t,k}|\mathcal{F}_{t,0}\right]$ is $\gamma_{t,k}(1 - \mu - \Lambda_{t,k+1})$; and the term in front of $\mathbb{E}\left[\mathcal{D}_{t,k}|\mathcal{F}_{t,0}\right]$ is $\gamma_{t,k}^{-1}(1 - 1/(4\mu) - \Lambda_{t,k+1})$. By symmetry, we choose $\mu = 1/2$ so that $\mu = 1/(4\mu)$. This yields

$$\begin{aligned}
&\mathbb{E}\left[\mathrm{W}(\widehat{S}_{t,k_t^{\mathrm{in}}}) + g(\widehat{S}_{t,k_t^{\mathrm{in}}})|\mathcal{F}_{t,0}\right] \\
&\quad\le \mathbb{E}\left[\mathrm{W}(\widehat{S}_{t,0}) + g(\widehat{S}_{t,0})|\mathcal{F}_{t,0}\right] \\
&\quad - \sum_{k=1}^{k_t^{\mathrm{in}}}\gamma_{t,k}\left(\frac{1}{2} - \Lambda_{t,k+1}\right)\mathbb{E}\left[\Delta^\star_{t,k}|\mathcal{F}_{t,0}\right] \\
&\quad - \sum_{k=1}^{k_t^{\mathrm{in}}}\frac{1}{\gamma_{t,k}}\left(\frac{1}{2} - \Lambda_{t,k+1}\right)\mathbb{E}\left[\mathcal{D}_{t,k}|\mathcal{F}_{t,0}\right] \\
&\quad + \gamma_{t,0} v_{\max} k_t^{\mathrm{in}}\|\mathcal{E}_t\|^2 \\
&\quad + v_{\max}\sum_{\ell=1}^{k_t^{\mathrm{in}}}(k_t^{\mathrm{in}} - \ell + 1)\gamma_{t,\ell}\,\mathcal{U}_{t,\ell} .
\end{aligned}$$

We now sum for $t = 1$ to $t = k^{\mathrm{out}}$ and compute the expectation. This yields, by using that $\widehat{S}_{t+1,0} = \widehat{S}_{t,k^{\mathrm{in}}}$,

$$\begin{aligned}
&\sum_{t=1}^{k^{\mathrm{out}}}\sum_{k=1}^{k_t^{\mathrm{in}}}\gamma_{t,k}\left(\frac{1}{2} - \Lambda_{t,k+1}\right)\mathbb{E}\left[\Delta^\star_{t,k}\right] \\
&+ \sum_{t=1}^{k^{\mathrm{out}}}\sum_{k=1}^{k_t^{\mathrm{in}}}\frac{1}{\gamma_{t,k}}\left(\frac{1}{2} - \Lambda_{t,k+1}\right)\mathbb{E}\left[\mathcal{D}_{t,k}\right] \\
&\le \quad \mathbb{E}\left[\mathrm{W}(\widehat{S}_{t,0}) + g(\widehat{S}_{t,0})\right] \\
&\quad - \mathbb{E}\left[\mathrm{W}(\widehat{S}_{k^{\mathrm{out}},k_{k^{\mathrm{out}}}^{\mathrm{in}}}) + g(\widehat{S}_{k^{\mathrm{out}},k_{k^{\mathrm{out}}}^{\mathrm{in}}})\right] \\
&\quad + v_{\max}\sum_{t=1}^{k^{\mathrm{out}}}\gamma_{t,0} k_t^{\mathrm{in}}\mathbb{E}\left[\|\mathcal{E}_t\|^2\right] \\
&\quad + v_{\max}\sum_{t=1}^{k^{\mathrm{out}}}\sum_{\ell=1}^{k_t^{\mathrm{in}}}\left(k_t^{\mathrm{in}} - \ell + 1\right)\gamma_{t,\ell}\,\mathcal{U}_{t,\ell} .
\end{aligned}$$

The proof is concluded upon noting that $\mathbb{E}\left[\mathrm{W}(\widehat{S}_{k^{\mathrm{out}},k_{k^{\mathrm{out}}}^{\mathrm{in}}}) + g(\widehat{S}_{k^{\mathrm{out}},k_{k^{\mathrm{out}}}^{\mathrm{in}}})\right] \ge \min_{\mathcal{S}}(W + g)$.

### 7.6 Proof of Corollary 4.3

Since $\mathcal{U}_{t,k} = 0$, we have $C_b = C_{vb} = 0$. In addition, $k_t^{\rm in} = k^{\rm in}$ for any $t$. Therefore, we can consider a constant stepsize sequence $\gamma_{t,k} = \gamma_\star$ where $\gamma_\star$ satisfies (see (15) and (16))

$$\gamma_\star \frac{L_{\dot{\rm W}}}{v_{\min}} + \gamma_\star^2 \frac{2v_{\max}}{v_{\min}} L^2 \frac{k^{\rm in}}{\mathsf{b}} \in (0, 1/2) \ .$$

This condition is satisfied by choosing

$$\frac{k^{\rm in}}{\mathsf{b}} \stackrel{\rm def}{=} \frac{1}{v_{\min} v_{\max}} \frac{L_{\dot{\rm W}}^2}{L^2} \ ,$$
$$\gamma_\star \stackrel{\rm def}{=} \frac{1}{4v_{\max}} \frac{L_{\dot{\rm W}}}{L^2} \frac{\mathsf{b}}{k^{\rm in}} = \frac{v_{\min}}{4L_{\dot{\rm W}}} \ .$$

Such a choice implies that $\inf_{t,k}(1/2 - \Lambda_{t,k}) = 1/2 - 3/8 = 1/8$. Since $\mathcal{E}_t = \mathcal{U}_{t,k} = 0$, we obtain from Corollary 4.2 that

$$\mathbb{E}\left[\Delta_{\tau,K}^\star + \mathcal{D}_{\tau,K}^\star\right]$$
$$\leq \frac{32\, L_{\dot{\rm W}}}{v_{\min}} \frac{\left(\mathbb{E}\left[{\rm W}(\widehat{S}_{1,0}) + g(\widehat{S}_{1,0})\right] - \min_{\mathcal{S}}\left({\rm W} + g\right)\right)}{k^{\rm out} k^{\rm in}} \ .$$

The $\epsilon$-approximate stationary condition is satisfied by choosing $k^{\rm out} k^{\rm in} = O(L_{\dot{\rm W}}/(v_{\min}\,\epsilon))$. The number of calls to the proximal operator is $k^{\rm out} k^{\rm in}$ so that $\mathcal{K}_{\rm prox} = O(L_{\dot{\rm W}}/(v_{\min}\,\epsilon))$. Finally, we have $\mathsf{b}_t' = n$ so that the number of calls to one of the $\bar{\mathsf{h}}_i$'s is $k^{\rm out}\, n + 2k^{\rm out} k^{\rm in} \mathsf{b}$. We can choose $\mathsf{b} = O\left(\sqrt{n}\sqrt{v_{\min} v_{\max}} L / L_{\dot{\rm W}}\right)$. This yields $k^{\rm out} = O(L\sqrt{v_{\max}}/(\sqrt{v_{\min}}\epsilon\sqrt{n}))$, and $\mathcal{K}_{\bar{\mathsf{h}}} = O(\sqrt{v_{\max}} L \sqrt{n}/(\epsilon\sqrt{v_{\min}}))$.

### 7.7 Cost of the approximation on the $\bar{\mathsf{h}}_i$'s

Following the rates obtained in Corollary 4.3, let us set $k_t^{\rm in} = O(\sqrt{n})$, $\mathsf{b} = O(\sqrt{n})$ and $k^{\rm out} = O(1/(\sqrt{n}\epsilon))$ and let us show that we can define random approximations $\delta_{t,0,i}$ and $\delta_{t,k+1,i}$ such that the $\epsilon$-approximate stationarity condition is satisfied.

**On the term $\mathbb{E}\left[\|\mathcal{E}_\tau\|^2\right]$.** We write $\mathbb{E}\left[\|\mathcal{E}_\tau\|^2\right] = (k^{\rm out})^{-1} \sum_{t=1}^{k^{\rm out}} \mathbb{E}\left[\|\mathcal{E}_t\|^2\right]$ and

$$\frac{1}{k^{\rm out}} \sum_{t=1}^{k^{\rm out}} \frac{\epsilon^{1-\mathsf{a}'}}{\sqrt{n}^{\mathsf{a}'} t^{\mathsf{a}'}} = \epsilon\, O(1) \ .$$

Let us compute the associated Monte Carlo complexity in the case $\mathcal{E}_t = n^{-1} \sum_{i=1}^n \{\delta_{t,0,i} - \bar{\mathsf{h}}_i(\widehat{S}_{t,0})\}$ and $\delta_{t,0,i}$ is equal to a Monte Carlo sum with $m_{t,0}$ i.i.d. samples. Then $\mathbb{E}\left[\|\mathcal{E}_t\|^2\right] = n^{-1} m_{t,0}^{-1} O(1)$. It is equal to $O(\epsilon^{1-\mathsf{a}'}/(\sqrt{n}t)^{\mathsf{a}'})$ when $m_{t,0} = O(n^{\mathsf{a}'/2-1} t^{\mathsf{a}'}/\epsilon^{1-\mathsf{a}'})$. Therefore, the Monte Carlo cost is

$$\sum_{t=1}^{k^{\rm out}} n m_{t,0} = O\left(\frac{1}{\sqrt{n}\epsilon^2}\right) \ .$$

**On the term $\mathbb{E}\left[\left(k^{\rm in} - K + 1\right)\mathcal{U}_{\tau,K}\right]$.** This term is upper bounded by $k^{\rm in}\,\mathbb{E}\left[\mathcal{U}_{\tau,K}\right]$ and we write

$$k^{\rm in}\,\mathbb{E}\left[\mathcal{U}_{\tau,K}\right] \leq \frac{1}{k^{\rm out}} \sum_{t=1}^{k^{\rm out}} \sum_{k=1}^{k^{\rm in}} O\left(\frac{1}{\mathsf{b}\, M_{t,k+1}}\right) \ .$$

The RHS is $O(\epsilon)$. The associated Monte Carlo complexity is

$$2\mathsf{b} \sum_{t=1}^{k^{\rm out}} \sum_{k=1}^{k^{\rm in}} M_{t,k} = O\left(\frac{\sqrt{n}}{\epsilon^2}\right) \ ,$$

whatever $\mathsf{a}, \bar{\mathsf{a}} \in [0, 1)$.

**Acknowledgments.** This work was partly supported by the Fondation Simone et Cino del Duca, under the program OpSiMorE; and by the french Agence Nationale de la Recherche (ANR) under the program ANR-19-CE23 MASDOL.

## Appendix A The condition A 4 in the Monte Carlo case

Following the framework detailed in Section 3.4, let us assume that *(i)* the intractable quantities $\mathsf{h}_i(\widehat{S}_{t,k}, B_{t,k+1})$ and $\mathsf{h}_i(\widehat{S}_{t,k-1}, B_{t,k})$ are of the form

$$\mathsf{h}_i(s, B) = \int_{\mathsf{Z}} H_\vartheta(z) \pi_\vartheta(\mathrm{d}z) \ , \qquad \text{(A1)}$$

where $\vartheta \stackrel{\rm def}{=} (s, i, B)$; and *(ii)* these integrals are approximated by a Monte Carlo sum: set $\vartheta_{t,\ell+1,i} \stackrel{\rm def}{=} (\widehat{S}_{t,\ell}, i, B_{t,\ell+1})$ and

$$\delta_{t,k+1,i} \stackrel{\text{def}}{=} \frac{1}{m_{t,k+1}} \sum_{r=1}^{m_{t,k+1}} \left\{ H_{\vartheta_{t,k+1,i}}(Z_r^{\vartheta_{t,k+1,i}}) - H_{\vartheta_{t,k,i}}(Z_r^{\vartheta_{t,k,i}}) \right\} , \quad \text{(A2)}$$

where, conditionally to $\widehat{S}_{t,k-1}$, $\widehat{S}_{t,k}$, $B_{t,k}$ and $B_{t,k+1}$, the samples $\{Z_r^{\vartheta_{t,\ell,i}}, r \geq 1\}$ are a Markov chain with unique stationary distribution $\pi_{\vartheta_{t,\ell,i}}(\mathrm{d}z)$; $\ell \in \{k, k+1\}$. Below, we show that A4 is verified when the Markov chain is ergodic enough. Let us start with introducing few notations from the Markov chain theory (see e.g. Meyn and Tweedie (1993)).

Let $P$ be a transition kernel onto the measurable set $(\mathsf{Z}, \mathcal{Z})$ and $\lambda, \pi$ be probability measures on $(\mathsf{Z}, \mathcal{Z})$. For a measurable function $\xi : \mathsf{Z} \to [0, +\infty)$, define

$$\pi(\xi) \stackrel{\text{def}}{=} \int_{\mathsf{Z}} \xi(z)\, \pi(\mathrm{d}z) .$$

For any $r \in \mathbb{N}$, the $r$-iterated transition kernel $P^r$ is defined by induction:

$$P^{r+1}(z, A) \stackrel{\text{def}}{=} \int_{\mathsf{Z}} P^r(z, \mathrm{d}y)\, P(y, A) = \int_{\mathsf{Z}} P(z, \mathrm{d}y)\, P^r(y, A) ,$$

for all $z \in \mathsf{Z}, A \in \mathcal{Z}$; by convention, $P^0(z, A) \stackrel{\text{def}}{=} \chi_A(z)$ the $\{0,1\}$-valued indicator function and $P^0(z, A) = \delta_z(A)$, the Dirac mass at zero. Given a probability measure $\lambda$ on $(\mathsf{Z}, \mathcal{Z})$, $\lambda P$ stands for the probability measure on $(\mathsf{Z}, \mathcal{Z})$ given by

$$\lambda P(A) \stackrel{\text{def}}{=} \int_{\mathsf{Z}} \lambda(\mathrm{d}y) P(y, A) , \qquad \forall A \in \mathcal{Z} .$$

For a function $U : \mathsf{Z} \to [1, +\infty)$ such that $\lambda P^r(U) + \pi(U) < +\infty$, define the $U$-norm of a measurable function $\xi : \mathsf{Z} \to \mathbb{R}^q$

$$\|\xi\|_U \stackrel{\text{def}}{=} \sup_{\mathsf{Z}} \frac{\|\xi\|}{U} ;$$

and the $U$-norm of the signed measure $\lambda P^r - \pi$ by

$$\|\lambda P^r - \pi\|_U \stackrel{\text{def}}{=} \sup_{\xi : \|\xi\|_U \leq 1} \|\lambda P^r(\xi) - \pi(\xi)\| .$$

Let us go back to sufficient conditions for verifying A4. Denote by $P_\vartheta$ a Markov transition kernel with invariant distribution $\pi_\vartheta(\mathrm{d}z)$: at iteration $(t, k+1)$, conditionally to $(\widehat{S}_{t,k-1}, \widehat{S}_{t,k}, B_{t,k}, B_{t,k+1})$, the chains $\{Z_r^{\vartheta_{t,k}}, r \geq 0\}$ and $\{Z_r^{\vartheta_{t,k+1}}, r \geq 0\}$ are Markov chains with transition kernels $P_{\vartheta_{t,k}}$ and $P_{\vartheta_{t,k+1}}$ respectively. They have the same initial value $\lambda$. Assume

**A5.** *1. There exists a measurable function $U : \mathsf{Z} \to [1, +\infty)$ such that*

$$H_\star \stackrel{\text{def}}{=} \sup_{(s,i,B) \in \mathcal{S} \times [n]^\star \times \mathcal{P}_+^q} \|H_\vartheta\|_U < +\infty ,$$

*where $H_\vartheta$ is defined by (A1).*

*2. There exist a function $\rho : \mathbb{N} \to [0, 1]$ and a positive constant $C_{\mathsf{MC}}$ such that for any $r \in \mathbb{N}$,*

$$\sup_{\vartheta \in \mathcal{S} \times [n]^\star \times \mathcal{P}_+^q} \|\lambda P_\vartheta^r - \pi_\vartheta\|_U \leq C_{\mathsf{MC}}\, \rho(r) .$$

*In addition, $\sum_{r \geq 1} \rho(r) < +\infty$.*

*3. Let $\vartheta \in \mathcal{S} \times [n]^\star \times \mathcal{P}_+^q$. Let $\{Z_r^\vartheta, r \geq 1\}$ be a Markov chain with transition kernel $P_\vartheta$ and initial distribution $\lambda$. There exists a positive constant $C'_{\mathsf{MC}}$ such that for any $\vartheta \in \mathcal{S} \times [n]^\star \times \mathcal{P}_+^q$ and $m' \in \mathbb{N}_\star$,*

$$\mathbb{E}\left[ \| \sum_{r=1}^{m'} \{H_\vartheta(Z_r^\vartheta) - \pi_\vartheta(H_\vartheta)\} \|^2 \right] \leq H_\star^2\, C'_{\mathsf{MC}}\, m' .$$

A5-item 2 is a uniform-in-$s$ ergodicity condition. Sufficient conditions for it are provided in (Fort et al, 2011, Lemma 2.3.) in the case of a geometric rate $\rho(r) = \kappa^r$ for some $\kappa \in (0, 1)$. By adapting (Andrieu et al, 2015, Theorem 1), similar conditions can be obtained in the case of a subgeometric rate $\rho(r)$. Sufficient conditions for A5-item 3 can be obtained from a trivial adaptation of (Fort and Moulines, 2003, Proposition 12).

We prove the following result.

**Proposition A.1** *Assume A 5. Let $\delta_{t,k+1,i}$ be given by (A2), where conditionally to $(\widehat{S}_{t,k-1}, \widehat{S}_{t,k}, B_{t,k}, B_{t,k+1})$, $\{Z_r^{\vartheta_{t,\ell,i}}, r \geq 0\}$ is a Markov chain with transition kernel $P_{\vartheta_{t,\ell,i}}$ and initial distribution $\lambda$, for $\ell \in \{k, k+1\}$. Then A 4 is verified with $m_{t,k+1} = M_{t,k+1} = \bar{M}_{t,k+1} \leftarrow m_{t,k+1}$,*

$C_b = C_{vb} \stackrel{\text{def}}{=} 2\,H_\star\,C_{\mathsf{MC}}\,\sum_{r\geq 1}\rho(r)$ *and* $C_v \stackrel{\text{def}}{=} 2H_\star^2\,C'_{\mathsf{MC}}$.

*Proof* We will use the notations

$$P_{\ell,i} \stackrel{\text{def}}{=} P_{\vartheta_{t,\ell+1,i}},\ \pi_{\ell,i} \stackrel{\text{def}}{=} \pi_{\vartheta_{t,\ell+1,i}},\ H_{\ell,i} \stackrel{\text{def}}{=} H_{\vartheta_{t,\ell+1,i}}\ .$$

- **Expression of $\mu_{t,k+1}$.** We have

$$\mu_{t,k+1,i} \stackrel{\text{def}}{=} \frac{1}{m_{t,k+1}}\sum_{r=1}^{m_{t,k+1}}\left(\lambda P_{k,i}^r H_{k,i} - \pi_{k,i}(H_{k,i})\right)$$
$$-\frac{1}{m_{t,k+1}}\sum_{r=1}^{m_{t,k+1}}\left(\lambda P_{k-1,i}^r H_{k-1,i} - \pi_{k-1,i}(H_{k-1,i})\right)\ .$$

- **The condition A4-Item 2.** By A5 and since $\widehat{S}_\bullet \in \mathcal{S}$, we write

$$\sup_{k,i}\|\lambda P_{k,i}^r H_{k,i} - \pi_{k,i}(H_{k,i})\| \leq H_\star\,C_{\mathsf{MC}}\,\rho(r)\ .$$

This implies that

$$\|\mu_{t,k+1,i}\| \leq 2\,H_\star\,C_{\mathsf{MC}}\,\frac{1}{m_{t,k+1}}\sum_{r=1}^{m_{t,k+1}}\rho(r)\ .$$

Since $\sum_r \rho(r) < \infty$, the RHS is of the form $C_b/m_{t,k+1}$ with $C_b \stackrel{\text{def}}{=} 2\,H_\star\,C_{\mathsf{MC}}\,\sum_r \rho(r)$.

- **The condition A4-Item 3.** We write

$$\sigma^2_{t,k+1,i} \leq \mathbb{E}\left[\|\xi_{t,k+1,i}\|^2|\mathcal{P}_{t,k+1/2}\right]\ .$$

Then, we have

$$\mathbb{E}\left[\|\xi_{t,k+1,i}\|^2|\mathcal{P}_{t,k+1/2}\right]$$
$$\leq 2\sup_{s,i}\mathbb{E}\left[\|\frac{1}{m_{t,k+1}}\sum_{r=1}^{m_{t,k+1}}H_{s,i}(Z_r^{s,i}) - \pi_{s,i}(H_{s,i})\|^2\right]$$

and the RHS is upper bounded by $2H_\star^2\,C'_{\mathsf{MC}}/m_{t,k+1}$ by A5-Item 3.

We also have

$$\frac{1}{n}\sum_{i=1}^n\|\mu_{t,k+1,i} - \frac{1}{n}\sum_{j=1}^n\mu_{t,k+1,j}\|^2 \leq \frac{1}{n}\sum_{i=1}^n\|\mu_{t,k+1,i}\|^2.$$

From the upper bound on $\|\mu_{t,k+1,i}\|$ above, we have

$$\|\mu_{t,k+1,i}\|^2 \leq \frac{C_b^2}{m_{t,k+1}^2}\ .$$

This concludes the proof. □

# Appendix B Supplementary materials for Section 5

## B.1 The penalized log-likelihood criterion

The observations are assumed independent, so the log-likelihood is given by

$$\theta \mapsto \sum_{i=1}^n \log\int_{\mathbb{R}^d}(1+\exp(-y_i\,\langle X_i, z_i\rangle))^{-1}$$
$$\times\,\frac{1}{\sqrt{2\pi}^d\,\sigma^d}\exp\left(-(2\sigma^2)^{-1}\|z_i-\theta\|^2\right)\mathrm{d}z_i\ .$$

The penalty term is $-n\tau\|\theta\|^2$.

**Lemma B.1** *The sum of the log-likelihood and the penalty term is equal to*

$$-\frac{n}{2}\log(2\pi\sigma^2) - \frac{1}{2\sigma^2}\theta^\top\sum_{i=1}^n\frac{X_iX_i^\top}{\|X_i\|^2}\theta - n\tau\|\theta\|^2$$
$$+\sum_{i=1}^n\log\int_{\mathbb{R}}\frac{\exp\left(x\,\langle X_i,\theta\rangle/(\sigma^2\|X_i\|)\right)}{1+\exp(-y_i\|X_i\|x)}\exp(-\frac{x^2}{2\sigma^2})\mathrm{d}x\ .$$

*Proof* Let $i \in [n]^\star$. Define an orthogonal $d\times d$ matrix $Q$ with columns denoted by $(Q_1,\cdots,Q_d)$, such that $Q_1 \stackrel{\text{def}}{=} X_i/\|X_i\|$. We have $\langle z_i, X_i\rangle = \|X_i\|\,\langle Q_1, z_i\rangle$, $\langle z_i,\theta\rangle = \left\langle Q^\top z_i, Q^\top\theta\right\rangle$ and $\|z_i\|^2 = \|Q^\top z_i\|^2$. This implies that

$$-\,y_i\,\langle z_i, X_i\rangle = -y_i\|X_i\|\,\langle Q_1, z_i\rangle$$
$$\|z_i-\theta\|^2 = \|\theta\|^2 + \|Q^Tz_i\|^2 - 2\left\langle Q^\top z_i, Q^\top\theta\right\rangle$$

so that the log-likelihood of the observation $Y_i$ is (up to the additive constant $C_1 \stackrel{\text{def}}{=} -d\ln\sigma - (d/2)\ln(2\pi)$)

$$y_i \mapsto -\frac{\|\theta\|^2}{2\sigma^2} + \log\int_{\mathbb{R}^d}(1+\exp(-y_i\|X_i\|\,\langle Q_1, z_i\rangle))^{-1}$$
$$\times\exp\left(-(2\sigma^2)^{-1}\left(\|Q^Tz_i\|^2 - 2\left\langle Q^\top z_i, Q^\top\theta\right\rangle\right)\right)\mathrm{d}z_i\ .$$

By a change of variable $v = (v_1,\cdots,v_q) \leftarrow Q^\top z_i$, the logarithm of the integral is equal to

$$\log\int_{\mathbb{R}^d}\frac{\exp\left(-(2\sigma^2)^{-1}\left(\|v\|^2 - 2\left\langle v, Q^\top\theta\right\rangle\right)\right)}{1+\exp(-y_i\|X_i\|v_1)}\mathrm{d}v$$
$$= \log\int_{\mathbb{R}}\frac{\exp\left(-(2\sigma^2)^{-1}\left(v_1^2 - 2v_1\,\langle Q_1,\theta\rangle\right)\right)}{1+\exp(-y_i\|X_i\|v_1)}\mathrm{d}v_1$$

$$+\sum_{u=2}^{d}\log\int_{\mathbb{R}}\exp\left(-\frac{1}{2\sigma^2}\{v_u^2-2v_u\,\langle Q_u,\theta\rangle\}\right)\,\mathrm{d}v_u\;.$$

The last $(d-1)$ integrals have a closed form. Observe indeed that $v_u^2-2v_u\,\langle Q_u,\theta\rangle=(v_u-\langle Q_u,\theta\rangle)^2-(\langle Q_u,\theta\rangle)^2$ so that up to the additive constant $C_2\stackrel{\text{def}}{=}(d-1)\{\log(2\pi)/2+\log\sigma\}$

$$\begin{aligned}&\sum_{u=2}^{d}\log\int_{\mathbb{R}}\exp\left(-\frac{1}{2\sigma^2}\{v_u^2-2v_u\,\langle Q_u,\theta\rangle\}\right)\,\mathrm{d}v_u\\&=\sum_{u=2}^{d}\frac{(\langle Q_u,\theta\rangle)^2}{2\sigma^2}=\frac{\|\theta\|^2-(\langle Q_1,\theta\rangle)^2}{2\sigma^2}\\&=\frac{\|\theta\|^2-(\langle X_i,\theta\rangle)^2/\|X_i\|^2}{2\sigma^2}\;.\end{aligned}$$

This concludes the proof; the constant (w.r.t. to $\theta$) is equal to $C_1+C_2$. □

## B.2 Proof of Lemma 5.1

The criterion $F$ is equal to $-\mathcal{L}(\theta)-\log(2\pi\sigma^2)/2$, where $\mathcal{L}(\theta)$ is the normalized penalized log-likelihood.

The likelihood is the product of probabilities, taking values in $(0,1)$; therefore, its logarithm is negative. The penalized log-likelihood is upper bounded $-\mathrm{pen}(\theta)=-n\tau\|\theta\|^2$. The normalized penalized log-likelihood is upper bounded $-\tau\|\theta\|^2$. Therefore the criterion is lower bounded by $\tau\|\theta\|^2-\ln(2\pi\sigma^2)/2$.

On the other hand, the minimum of the criterion is smaller than the value of the criterion at $\theta=0$. Let us show that this value is $(\ln 4)/n-\ln(2\pi\sigma^2)/2$. This will imply that the minimizers of the criterion are in the set $\{\theta\in\mathbb{R}^d:\tau\|\theta\|^2\le\ln 4\}$ and conclude the proof.

We have $\mathrm{pen}(0)=0$. Let us lower bound the likelihood of an observation $Y_i=+1$ at $\theta=0$. The likelihood is equal to

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

By using the same change of variable than in the proof of Lemma B.1, it is equal to

$$\frac{1}{\sqrt{2\pi}\sigma}\int_{\mathbb{R}}\frac{\exp\left(-x^2/(2\sigma^2)\right)}{1+\exp(-\|X_i\|x)}\mathrm{d}x\left(\frac{1}{\sqrt{2\pi}\sigma}\int_{\mathbb{R}}\exp\left(-x^2/(2\sigma^2)\right)\mathrm{d}x\right)^{d-1}\;,$$

and is lower bounded by (note that the $(d-1)$ identical integrals are equal to one)

$$\frac{1}{\sqrt{2\pi}\sigma}\int_{\mathbb{R}^+}\frac{\exp\left(-x^2/(2\sigma^2)\right)}{1+\exp(-\|X_i\|x)}\mathrm{d}x\;,$$

which is in turn lower bounded by $1/4$ since $1+\exp(-\|X_i\|x)\le 2$ for all $x\ge 0$.

The proof for the case $Y_i=-1$ is on the same lines and is omitted.

This implies that the likelihood of the $n$ variables is lower bounded by $1/4^n$; the normalized log-likelihood is lower bounded by $-\ln 4$; the criterion is upper bounded by $\ln 4-\ln(2\pi\sigma^2)/2$.

## B.3 The optimization problem seen as an EM

We established that for any $\theta\in\mathbb{R}^d$, $\nabla F(\theta)=n^{-1}\sum_{i=1}^n G_i(\theta)$ where

$$G_i(\theta)\stackrel{\text{def}}{=}2U\theta-\frac{X_i}{\sigma^2\,\|X_i\|}\int_{\mathbb{R}}z\;\pi_{\theta,i}(z)\mathrm{d}z\;,$$

and $\pi_{\theta,i}(z)$ is the probability density proportional to (20).

From the expressions of $\phi,\psi$ and $S(Y_i,z)$, we obtain that $\mathsf{T}$, defined by Proposition 2.1, is given by $\mathsf{T}(s)\stackrel{\text{def}}{=}U^{-1}s/2$ for any $s\in\mathbb{R}^d$. This implies that for any $s\in\mathbb{R}^d$,

$$\begin{aligned}\mathsf{h}_i(s,B)&\stackrel{\text{def}}{=}\int_{\mathbb{R}}S(Y_i,z)\,\pi_{\mathsf{T}(s),i}(z)\,\mathrm{d}z-s\\&=\frac{X_i}{\sigma^2\,\|X_i\|}\int_{\mathbb{R}}z\,\pi_{Bs,i}(z)\,\mathrm{d}z-s\;.\end{aligned}$$

For any $s\in\mathbb{R}^d$, let us find the matrix $\mathsf{B}(s)$ satisfying $\nabla(F\circ\mathsf{T})(s)=-n^{-1}\sum_{i=1}^n\mathsf{B}(s)\mathsf{h}_i(s)$ (see (6)). We have $\nabla(F\circ\mathsf{T})=\nabla F(B\cdot)=B^\top(\nabla F)(B\cdot)=B\,(\nabla F)(B\cdot)$. This yields

$$\begin{aligned}&\nabla(F\circ\mathsf{T})(s)=B\frac{1}{n}\sum_{i=1}^n G_i(Bs)\\&=B\frac{1}{n}\sum_{i=1}^n\left(2U\theta-\frac{X_i}{\sigma^2\,\|X_i\|}\int_{\mathbb{R}}z\;\pi_{Bs,i}(z)\mathrm{d}z\right)\\&=Bs-B\frac{1}{n}\sum_{i=1}^n\frac{X_i}{\sigma^2\,\|X_i\|}\int_{\mathbb{R}}z\;\pi_{Bs,i}(z)\mathrm{d}z\end{aligned}$$

$$= -B\frac{1}{n}\sum_{i=1}^{n} \mathsf{h}_i(s) \ .$$

This yields $\mathsf{B}(s) \stackrel{\text{def}}{=} B$ for any $s \in \mathbb{R}^d$.

## B.4 Proof of Lemma 5.2

Let $i \in [n]^\star$ and $\theta \in \mathbb{R}^d$. **Step 1.** By using

$$\begin{aligned} &-z^2/(2\sigma^2) + z\left\langle X_i, \theta\right\rangle/(\sigma^2\|X_i\|) \\ &\quad = -(z - \left\langle X_i, \theta\right\rangle/\|X_i\|)^2/(2\sigma^2) \\ &\qquad + (\left\langle X_i, \theta\right\rangle)^2/(2\sigma^2\|X_i\|^2) \ , \end{aligned}$$

we write

$$\pi_{\theta,i}(z) = \frac{\exp\left(-\left(z - \left\langle X_i, \theta\right\rangle/\|X_i\|\right)^2/(2\sigma^2)\right)}{Z_{\theta,i}\ \ 1+\exp(-y_i\|X_i\|z)}$$

where $Z_{\theta,i}$ is the normalizing constant. Second, we use $z\,\pi_{\theta,i}(z) = (z-a_i)\,\pi_{\theta,i}(z) + a_i\,\pi_{\theta,i}(z)$ with $a_i \leftarrow \left\langle X_i, \theta\right\rangle/\|X_i\|$ and since $\int_{\mathbb{R}} \pi_{\theta,i}(z)\mathrm{d}z = 1$, we obtain

$$\mathcal{I}_i(\theta) = \frac{\left\langle X_i, \theta\right\rangle}{\|X_i\|} + \int_{\mathbb{R}} \left(z - \frac{\left\langle X_i, \theta\right\rangle}{\|X_i\|}\right) \pi_{\theta,i}(z)\,\mathrm{d}z \ .$$

Finally, the integral in the RHS being of the form

$$\sigma^2 \int_{\mathbb{R}} \frac{f'(z)}{Z_{\theta,i}\ \ 1+\exp(-y_i\|X_i\|z)}\mathrm{d}z$$

with

$$f(z) \stackrel{\text{def}}{=} -\exp\left(-\left(z - \left\langle X_i, \theta\right\rangle/\|X_i\|\right)^2/(2\sigma^2)\right) \ ,$$

we use an integration by parts. Upon noting that the derivative of $z \mapsto 1/(1+\exp(-y_i\|X_i\|z))$ is

$$y_i\|X_i\|\frac{\exp(-y_i\|X_i\|z)}{\left(1+\exp(-y_i\|X_i\|z)\right)^2} \ ,$$

we write

$$\begin{aligned} &\int_{\mathbb{R}} \frac{f'(z)}{1+\exp(-y_i\|X_i\|z)}\mathrm{d}z \\ &= -y_i\|X_i\| \int f(z)\frac{\exp(-y_i\|X_i\|z)}{\left(1+\exp(-y_i\|X_i\|z)\right)^2}\mathrm{d}z \ . \end{aligned}$$

Therefore, the conclusion of this first step is

$$\begin{aligned} \mathcal{I}_i(\theta) &= \left\langle \frac{X_i}{\|X_i\|}, \theta \right\rangle \\ &\quad + y_i\|X_i\|\sigma^2 \int_{\mathbb{R}} \frac{\pi_{\theta,i}(z)}{1+\exp\left(y_i\|X_i\|z\right)}\,\mathrm{d}z \ . \end{aligned}$$

**Step 2.** This step is classical in the MCMC literature (see e.g. Choi and Hobert (2013) and references therein). We prove that for any $z \in \mathbb{R}$,

$$\pi_{\theta,i}(z) = \int_0^{+\infty} \bar{\pi}_{\theta,i}(z,\omega)\mathrm{d}\omega \ .$$

By (Polson et al, 2013, Theorem 1), it holds

$$\begin{aligned} \frac{1}{1+\exp\left(-y_i\|X_i\|z\right)} &= \frac{1}{2}\exp\left(y_i\|X_i\|z/2\right) \\ &\quad \times \int_0^{+\infty} \exp(-\omega\|X_i\|^2z^2/2)\mathsf{p}(\omega;\ 1)\mathrm{d}\omega \ , \end{aligned}$$

where $\mathsf{p}(\omega;\ b)\mathrm{d}\omega$ is a Polya-Gamma distribution with parameter $b$. This implies that $\pi_{\theta,i}(z)$ is equal to

$$\begin{aligned} &\exp\left(\frac{y_i\|X_i\|z}{2} - \frac{\left(z - \left\langle X_i, \theta\right\rangle/\|X_i\|\right)^2}{2\sigma^2}\right) \\ &\times \frac{1}{2Z_{\theta,i}} \int_0^{+\infty} \exp(-\omega\|X_i\|^2z^2/2)\mathsf{p}(\omega;\ 1)\mathrm{d}\omega \ . \end{aligned}$$

This concludes the proof.

## B.5 The assumption A4 is verified.

Define the Markov kernel with density

$$P_{s,i}(z;\ z') \stackrel{\text{def}}{=} \left(\int_0^{\infty} \pi_2(\omega|z;\ i)\,\pi_1(z'|\omega;\ s,i)\ \mathrm{d}\omega\right)$$

w.r.t. the Lebesgue measure on $\mathbb{R}$; here, $\pi_1(z'|\omega;\ s,i)$ is the density of a Gaussian distribution with expectation $\mathsf{m}_{s,i}(\omega)$ and variance $v_i(\omega)$ given by

$$\begin{aligned} v_i(\omega) &\stackrel{\text{def}}{=} \frac{\sigma^2}{1+\omega\sigma^2\|X_i\|^2} \ , \\ \mathsf{m}_{s,i}(\omega) &\stackrel{\text{def}}{=} v_i(\omega)\left(\frac{1}{\sigma^2}\left\langle \frac{X_i}{\|X_i\|}, Bs\right\rangle + \frac{1}{2}y_i\|X_i\|\right) \ ; \end{aligned}$$

and $\pi_2(\omega|z;\ i)$ is a Polya-Gamma distribution with parameter $(1, \|X_i\|z)$. The Gibbs kernel described by Lemma 5.2 and targeting the density

distribution $\bar{\pi}_{Bs,i}(z,\omega)\mathrm{d}z\mathrm{d}\omega$, produces a Markov chain $\{(Z_r^{s,i},\Omega_r^{s,i}), r\geq 0\}$ such that the marginal $\{Z_r^{s,i}, r\geq 0\}$ is a Markov chain with transition kernel $P_{s,i}(z;\ z')\,\mathrm{d}z'$. We apply the results of (Choi and Hobert, 2013, Proposition 3.1) with

| | |
|---|---|
| $y\in\{0,1\}$ | $y_i\in\{-1,1\}$ |
| $n$ | 1 |
| $\Omega(\omega)$ | $\omega$ |
| $X$ | $\|X_i\|$ |
| $B$ | $\sigma^2$ |
| $b$ | $\langle X_i, Bs\rangle/\|X_i\|$ |

**Table B1** [left] The notations of Choi and Hobert (2013). [right] the notations in this paper.

This yields

$$\int_A P_{s,i}(z;\ z')\mathrm{d}z' \geq \varepsilon \int_A \exp(-0.5(x-\mathsf{m}_\star)^2/v_\star^2)\,\mathrm{d}x \qquad \text{(B3)}$$

where

$$\varepsilon \stackrel{\text{def}}{=} \inf_{s\in\mathcal{S},i\in[n]^\star} \frac{\exp\left(-\frac{1}{4}-\frac{\{\mathsf{m}_{s,i}(1/2)\}^2\sigma^2\|X_i\|^2}{4\,v_i(1/2)}\right)}{2\sqrt{1+\sigma^2\|X_i\|^2/2}}$$

and $(\mathsf{m}_\star, v_\star)$ satisfy for any $x\in\mathbb{R}$,

$$\inf_{s\in\mathcal{S},i\in[n]^\star} \exp(-0.5\,(x-\mathsf{m}_{s,i}(1/2))^2/v_i(1/2)) \geq \exp(-0.5(x-\mathsf{m}_\star)^2/v_\star^2)\ .$$

**Lemma B.2** *Since $\mathcal{S}$ is bounded, then $\varepsilon>0$ and $\mathsf{m}_\star, v_\star$ exist in $\mathbb{R}\times(0,+\infty)$.*

The minorization condition (B3) implies that the kernel $P_{s,i}(z,z')\mathrm{d}z'$ is uniformly ergodic, uniformly in $s,i$ and $z$. By (Meyn and Tweedie, 1993, Theorem 16.0.2.) and (Fort and Moulines, 2003, Proposition 1), A 5-Item 2 and A 5-Item 3 are satisfied.

# Appendix C Detailed proofs

## C.1 Proof of (17)

Let $t\in[k^{\text{out}}]^\star$. The sequence given by $\gamma_{t,k+1}\stackrel{\text{def}}{=}\prod_{j=0}^k\left(1+\frac{2\,C_b}{m_{t,j+1}}\right)^{-1}\gamma_{t,0}$ for any $k\geq 0$, satisfies

$$\gamma_{t,k+1}\left(1+\frac{2C_b}{m_{t,k+1}}\right)\leq\gamma_{t,k}\ .$$

A sufficient condition for the property $\Lambda_{t,k+1}\in(0,1/2)$ to hold is $\mathsf{a}\gamma_{t,0}^2+\bar{\mathsf{a}}\gamma_{t,0}-1/2<0$ where

$$\bar{\mathsf{a}}\stackrel{\text{def}}{=}\frac{L_{\dot{\mathrm{W}}}}{v_{\min}}\ ,$$
$$\mathsf{a}\stackrel{\text{def}}{=}L^2\frac{2v_{\max}k_t^{\text{in}}}{v_{\min}\mathsf{b}}\left(1+\frac{2\,C_{vb}}{\sqrt{\mathsf{b}}\,\bar{M}_{t,k+1}}\right)\ .$$

The function $x\mapsto\mathsf{a}x^2+\bar{\mathsf{a}}x-1/2$ possesses two roots: one is positive and one is negative. The positive one is given by $(-\bar{\mathsf{a}}+\sqrt{\bar{\mathsf{a}}^2+2\mathsf{a}})/(2\mathsf{a})$; it is equal to (17).

## C.2 Proof of (18) and (19)

We write $\mathsf{S}_{t,0}-\mathsf{h}(\widehat{S}_{t,0})=U+V$ where

$$U\stackrel{\text{def}}{=}\frac{1}{\mathsf{b}_t'}\sum_{i\in\mathcal{B}_{t,0}}\left(\delta_{t,0,i}-\mathbb{E}\left[\delta_{t,0,i}|\widehat{S}_{t,0},\mathcal{B}_{t,0}\right]\right)\ ,$$
$$V\stackrel{\text{def}}{=}\frac{1}{\mathsf{b}_t'}\sum_{i\in\mathcal{B}_{t,0}}\mathbb{E}\left[\delta_{t,0,i}|\widehat{S}_{t,0},\mathcal{B}_{t,0}\right]-\mathsf{h}(\widehat{S}_{t,0})\ .$$

We have $\mathbb{E}\left[\|U+V\|^2\right]=\mathbb{E}\left[\|U\|^2\right]+\mathbb{E}\left[\|V\|^2\right]$ by definition of the conditional expectation. Since $\delta_{t,0,i}$ is an unbiased random approximation of $\mathsf{h}_i(\widehat{S}_{t,0})$, we have $\mathbb{E}\left[\delta_{t,0,i}|\widehat{S}_{t,0},\mathcal{B}_{t,0}\right]=\mathsf{h}_i(\widehat{S}_{t,0})$.

In the case $\mathcal{B}_{t,0}=\{1,\cdots,n\}$, then $V=0$. Therefore,

$$\begin{aligned}&\mathbb{E}\left[\|\mathsf{S}_{t,0}-\mathsf{h}(\widehat{S}_{t,0})\|^2\right]\\&=\frac{1}{n^2}\mathbb{E}\left[\|\sum_{i=1}^n\{\delta_{t,0,i}-\mathbb{E}\left[\delta_{t,0,i}|\widehat{S}_{t,0}\right]\}\|^2\right]\\&=\frac{1}{n^2}\sum_{i=1}^n\sigma_{t,0,i}^2\ ,\end{aligned}$$

where we used that the variables $\{\delta_{t,0,i}, i\in[n]^\star\}$ are independent conditionally to $\widehat{S}_{t,0}$, and with variance $\sigma_{t,0,i}^2$.

In the case $\mathcal{B}_{t,0}$ is a subset of $[n]^\star$ of cardinality $\mathsf{b}'_t$, then we write

$$\mathbb{E}\left[\|U\|^2\right] = \frac{1}{(\mathsf{b}'_t)^2}\mathbb{E}\left[\mathbb{E}\left[\|U\|^2|\mathcal{B}_{t,0}, \widehat{S}_{t,0}\right]\right]$$
$$= \frac{1}{\mathsf{b}'_t}\mathbb{E}\left[\frac{1}{\mathsf{b}'_t}\sum_{i\in\mathcal{B}_{t,0}}\sigma^2_{t,0,i}\right]$$

and we conclude by Lemma 7.1. Again from Lemma 7.1, we have

$$\mathbb{E}\left[\|V\|^2\right] \le \frac{1}{\mathsf{b}'_t\, n}\sum_{i=1}^n \|\mathsf{h}_i(\widehat{S}_{t,0}) - \mathsf{h}(\widehat{S}_{t,0})\|^2 ,$$

with an equality when $\mathcal{B}_{t,0}$ is sampled with replacement. This concludes the proof.

## C.3 Proof of Lemma 7.1

Set, for ease of notations,

$$\mathcal{B} \stackrel{\text{def}}{=} \mathcal{B}_{t,k}, \qquad \mathsf{h}_{\mathcal{B}} \stackrel{\text{def}}{=} \frac{1}{\mathsf{b}}\sum_{i\in\mathcal{B}}\mathsf{h}_i .$$

### C.3.1 Case with replacement

We write $\mathcal{B} = \{I_1, \cdots, I_{\mathsf{b}}\}$ where the r.v. $I_i$'s are independent, and uniformly distributed on $[n]^\star$.

• Then

$$\mathbb{E}\left[f_{\mathcal{B}}\right] = \frac{1}{\mathsf{b}}\sum_{\ell=1}^{\mathsf{b}}\mathbb{E}\left[f_{I_\ell}\right] = \mathbb{E}\left[f_{I_1}\right] = n^{-1}\sum_{i=1}^n f_i(u) .$$

• Set $\bar{f} \stackrel{\text{def}}{=} n^{-1}\sum_{i=1}^n f_i$. We have, by using that the r.v. $\{I_1, \cdots, I_{\mathsf{b}}\}$ are independent,

$$\mathbb{E}\left[\|f_{\mathcal{B}} - \bar{f}\|^2\right] = \mathbb{E}\left[\|\frac{1}{\mathsf{b}}\sum_{\ell=1}^{\mathsf{b}}\left(f_{I_\ell} - \bar{f}\right)\|^2\right]$$
$$= \frac{1}{\mathsf{b}^2}\sum_{\ell=1}^{\mathsf{b}}\mathbb{E}\left[\|f_{I_\ell} - \bar{f}\|^2\right]$$
$$= \frac{1}{\mathsf{b}}\mathbb{E}\left[\|f_{I_1} - \bar{f}\|^2\right] = \frac{1}{\mathsf{b}\,n}\sum_{i=1}^n \|f_i - \bar{f}\|^2 .$$

• Since the variance of the sum is the sum of the variance for independent r.v.

$$\mathbb{E}\left[\|\mathsf{h}_{\mathcal{B}}(u) - \mathsf{h}_{\mathcal{B}}(u') - \{\mathsf{h}(u) - \mathsf{h}(u')\}\|^2\right]$$
$$= \frac{1}{\mathsf{b}^2}\sum_{\ell=1}^{\mathsf{b}}\mathbb{E}\left[\|\mathsf{h}_{I_\ell}(u) - \mathsf{h}_{I_\ell}(u') - \mathsf{h}(u) + \mathsf{h}(u')\|^2\right] .$$

Then, since $I_\ell$ is uniformly distributed on $[n]^\star$,

$$\mathbb{E}\left[\|\mathsf{h}_{I_\ell}(u) - \mathsf{h}_{I_\ell}(u') - \mathsf{h}(u) + \mathsf{h}(u')\|^2\right]$$
$$= \frac{1}{n}\sum_{i=1}^n \mathbb{E}\left[\|\mathsf{h}_i(u) - \mathsf{h}_i(u')\|^2\right] - \|\mathsf{h}(u) - \mathsf{h}(u')\|^2$$
$$\le \|u - u'\|^2\frac{1}{n}\sum_{i=1}^n L_i^2 - \|\mathsf{h}(u) - \mathsf{h}(u')\|^2 . \tag{C4}$$

### C.3.2 Case without replacement

Set $\mathcal{B} = \{I_1, \cdots, I_{\mathsf{b}}\}$ and $\bar{f} \stackrel{\text{def}}{=} n^{-1}\sum_{i=1}^n f_i$. $I_1$ is a uniform random variable on $[n]^\star$ so that $\mathbb{E}\left[f_{I_1}\right] = \bar{f}$.

• Conditionally to $I_1$, $I_2$ is a uniform random variable on $[n]^\star \setminus \{I_1\}$. Therefore

$$\mathbb{E}\left[f_{I_2}\right] = \frac{1}{n-1}\left(\sum_{j=1}^n f_j - \mathbb{E}\left[f_{I_1}\right]\right)$$
$$= \frac{n}{n-1}\bar{f} - \frac{1}{n-1}\bar{f} = \bar{f} .$$

By induction, for any $\ell \ge 2$,

$$\mathbb{E}\left[f_{I_\ell}\right]$$
$$= \frac{1}{n-\ell+1}\left(\sum_{j=1}^n f_j - \sum_{r=1}^{\ell-1}\mathbb{E}\left[f_{I_r}\right]\right)$$
$$= \frac{n}{n-\ell+1}\bar{f} - \frac{\ell-1}{n-\ell+1}\bar{f} = \bar{f} .$$

As a conclusion, $\mathsf{b}^{-1}\sum_{\ell=1}^{\mathsf{b}}\mathbb{E}\left[f_{I_\ell}\right] = \bar{f}$.

• Let $u, u' \in \mathcal{S}$; set $\phi(I_\ell) \stackrel{\text{def}}{=} \mathsf{h}_{I_\ell}(u) - \mathsf{h}(u) - \mathsf{h}_{I_\ell}(u') + \mathsf{h}(u')$. Then $\mathbb{E}\left[\phi(I_\ell)\right] = 0$. First, we prove by induction that $\mathbb{E}\left[\|\phi(I_\ell)\|^2\right] = \mathbb{E}\left[\|\phi(I_1)\|^2\right]$. Upon noting that $I_1$ is a uniform random variable on $[n]^\star$ and by using the induction assumption,

$$(n-\ell+1)\mathbb{E}\left[\|\phi(I_\ell)\|^2\right]$$
$$= \left(\sum_{i=1}^n \|\phi(i)\|^2 - \mathbb{E}\left[\sum_{r=1}^{\ell-1}\|\phi(I_r)\|^2\right]\right)$$
$$= n\mathbb{E}\left[\|\phi(I_1)\|^2\right] - (\ell-1)\mathbb{E}\left[\|\phi(I_1)\|^2\right] ,$$

which concludes the induction. Second, let us prove that for any $\ell \geq 0$,

$$\mathbb{E}\left[\|\sum_{r=1}^{\ell+1}\phi(I_r)\|^2\right] \leq (\ell+1)\mathbb{E}\left[\|\phi(I_1)\|^2\right] . \quad \text{(C5)}$$

Since $\sum_{i=1}^{n}\phi(i) = n\mathbb{E}\left[\phi(I_1)\right] = 0$,

$$\begin{aligned}
&\mathbb{E}\left[\left\langle\sum_{p=1}^{\ell}\phi(I_p), \phi(I_{\ell+1})\right\rangle\right] \\
&= \mathbb{E}\left[\left\langle\sum_{p=1}^{\ell}\phi(I_p), \mathbb{E}\left[\phi(I_{\ell+1})\middle| I_1, \cdots, I_\ell\right]\right\rangle\right] \\
&= \frac{1}{n-\ell}\mathbb{E}\left[\left\langle\sum_{p=1}^{\ell}\phi(I_p), \sum_{i=1}^{n}\phi(i) - \sum_{p=1}^{\ell}\phi(I_p)\right\rangle\right] \\
&= -\frac{1}{n-\ell}\mathbb{E}\left[\|\sum_{p=1}^{\ell}\phi(I_p)\|^2\right] ,
\end{aligned}$$

so that

$$\begin{aligned}
&\mathbb{E}\left[\|\sum_{p=1}^{\ell+1}\phi(I_p)\|^2\right] \\
&= \left(1 - \frac{2}{n-\ell}\right)\mathbb{E}\left[\|\sum_{p=1}^{\ell}\phi(I_p)\|^2\right] + \mathbb{E}\left[\|\phi(I_{\ell+1})\|^2\right] \\
&\leq (\ell+1)\mathbb{E}\left[\|\phi(I_1)\|^2\right] .
\end{aligned}$$

The proof of the first bound follows from (C5) and (C4) since here again, $I_1$ is uniformly distributed on $[n]^\star$.

- The proof of the second bound is similar (change the definition of $\phi(I_\ell)$); it is omitted.

## References


Allen-Zhu Z (2018) Natasha 2: Faster Non-Convex Optimization Than SGD. In: Bengio S, Wallach H, Larochelle H, et al (eds) Advances in Neural Information Processing Systems, vol 31. Curran Associates, Inc.

Allen-Zhu Z, Hazan E (2016) Variance reduction for faster non-convex optimization. In: Balcan M, Weinberger K (eds) 33rd International Conference on Machine Learning, ICML 2016. International Machine Learning Society (IMLS), 33rd International Conference on Machine Learning, ICML 2016, pp 1093–1101

Andrieu C, Fort G, Vihola M (2015) Quantitative convergence rates for subgeometric markov chains. J Appl Probab 52(2):391–404. https://doi.org/10.1239/jap/1437658605

Atchadé Y, Fort G, Moulines E (2017) On Perturbed Proximal Gradient Algorithms. Journal of Machine Learning Research 18(10):1–33

Bauschke HH, Combettes PL (2011) Convex Analysis and Monotone Operator Theory in Hilbert Spaces, 1st edn. Springer Publishing Company, Incorporated, https://doi.org/10.1007/978-1-4419-9467-7

Beck A (2017) First-Order Methods in Optimization. Society for Industrial and Applied Mathematics, Philadelphia, PA, https://doi.org/10.1137/1.9781611974997

Becker S, Fadili J (2012) A quasi-Newton proximal splitting method. In: Pereira F, Burges C, Bottou L, et al (eds) Advances in Neural Information Processing Systems, vol 25. Curran Associates, Inc.

Becker S, Fadili J, Ochs P (2019) On quasi-newton forward-backward splitting: Proximal calculus and convergence. SIAM Journal on Optimization 29(4):2445–2481. https://doi.org/10.1137/18M1167152

Benveniste A, Métivier M, Priouret P (1990) Adaptive Algorithms and Stochastic Approximations. Springer Verlag, https://doi.org/https://doi.org/10.1007/978-3-642-75894-2

Bonettini S, Porta F, Ruggiero V, et al (2021) Variable metric techniques for forward–backward methods in imaging. Journal of Computational and Applied Mathematics 385:113,192. https://doi.org/https://doi.org/10.1016/j.cam.2020.113192

Borkar VS (2008) Stochastic approximation. Cambridge University Press, Cambridge; Hindustan Book Agency, New Delhi, https://doi.org/https://doi.org/10.1007/978-93-86279-38-5, a

dynamical systems viewpoint

Brown L (1986) Fundamentals of statistical exponential families : with applications in statistical decision theory. Lecture notes-monograph series Fundamentals of statistical exponential families, Institute of Mathematical Statistics, https://doi.org/10.1214/lnms/1215466757

Cappé O, Moulines E (2009) On-line Expectation Maximization algorithm for latent data models. J Roy Stat Soc B Met 71(3):593–613. https://doi.org/https://doi.org/10.1111/j.1467-9868.2009.00698.x

Celeux G, Diebolt J (1985) The SEM algorithm: a probabilistic teacher algorithm derived from the EM algorithm for the mixture problem. Computational Statistics Quarterly 2:73–82

Chen HG, Rockafellar R (1997) Convergence rates in forward-backward splitting. SIAM J Optim 7:421–444. https://doi.org/https://doi.org/10.1137/S1052623495290179

Chen J, Zhu J, Teh Y, et al (2018) Stochastic Expectation Maximization with Variance Reduction. In: Bengio S, Wallach H, Larochelle H, et al (eds) Advances in Neural Information Processing Systems 31. Curran Associates, Inc., p 7967–7977, https://doi.org/10.5555/3327757.3327893

Chen X, Liu S, Sun R, et al (2019) On the convergence of a class of adam-type algorithms for non-convex optimization. In: International Conference on Learning Representations

Choi H, Hobert J (2013) The Polya-Gamma Gibbs sampler for Bayesian logistic regression is uniformly ergodic. Electronic Journal of Statistics 7(none):2054 – 2064. https://doi.org/10.1214/13-EJS837

Chouzenoux E, Pesquet JC, Repetti A (2014) Variable Metric Forward–Backward Algorithm for Minimizing the Sum of a Differentiable Function and a Convex Function. Journal of Optimization Theory and Applications 162(1):107–132. https://doi.org/10.1007/s10957-013-0465-7

Combettes P, Pesquet J (2011) Proximal Splitting Methods in Signal Processing. In: Bauschke HH, Burachik RS, Combettes PL, et al (eds) Fixed-Point Algorithms for Inverse Problems in Science and Engineering. Springer Optimization and Its Applications, Springer, p 185–212, https://doi.org/10.1007/978-1-4419-9569-8

Combettes P, Vũ B (2014) Variable metric forward–backward splitting with applications to monotone inclusions in duality. Optimization 63(9):1289–1318. https://doi.org/10.1080/02331934.2012.733883

Combettes PL, Wajs VR (2005) Signal Recovery by Proximal Forward-Backward Splitting. Multiscale Modeling & Simulation 4(4):1168–1200. https://doi.org/10.1137/050626090

Defazio A, Bach F, Lacoste-Julien S (2014) Saga: A fast incremental gradient method with support for non-strongly convex composite objectives. In: Proceedings of the 27th International Conference on Neural Information Processing Systems - Volume 1. MIT Press, Cambridge, MA, USA, NIPS'14, p 1646–1654

Delyon B, Lavielle M, Moulines E (1999) Convergence of a Stochastic Approximation version of the EM algorithm. Ann Statist 27(1):94–128. https://doi.org/10.1214/aos/1018031103

Dempster A, Laird N, Rubin D (1977) Maximum Likelihood from Incomplete Data via the EM Algorithm. J Roy Stat Soc B Met 39(1):1–38

Devroye L (1986) Non-Uniform Random Variate Generation(originally published with. Springer-Verlag, https://doi.org/https://doi.org/10.1007/978-1-4613-8643-8

Eicke B (1992) Iteration methods for convexly constrained ill-posed problems in hilbert space. Numerical Functional Analysis and Optimization 13(5-6):413–429. https://doi.org/10.1080/01630569208816489

Everitt B (1984) An introduction to latent variable models. Chapman and Hall London ; New York, https://doi.org/https://doi.org/10.1007/978-94-009-5564-6

Fang C, Li C, Lin Z, et al (2018) Spider: Near-optimal non-convex optimization via stochastic path-integrated differential estimator. In: Bengio S, Wallach H, Larochelle H, et al (eds) Advances in Neural Information Processing Systems, vol 31. Curran Associates, Inc.

Fort G, Moulines E (2003) Convergence of the Monte Carlo Expectation Maximization for curved exponential families. Ann Statist 31(4):1220–1259

Fort G, Moulines E (2021) The Perturbed Prox-Preconditioned Spider Algorithm: Non-Asymptotic Convergence Bounds. In: 2021 IEEE Statistical Signal Processing Workshop (SSP), pp 96–100, https://doi.org/10.1109/SSP49050.2021.9513846

Fort G, Moulines E, Priouret P (2011) Convergence of adaptive and interacting markov chain monte carlo algorithms. Ann Statist 39(6):3262–3289

Fort G, Risser L, Atchadé Y, et al (2018) Stochastic fista algorithms: So fast ? In: 2018 IEEE Statistical Signal Processing Workshop (SSP), pp 796–800, https://doi.org/10.1109/SSP.2018.8450740

Fort G, Moulines E, Wai HT (2020) A Stochastic Path-Integrated Differential Estimator Expectation Maximization Algorithm. In: Proceedings of the 34th International Conference on Neural Information Processing Systems. Curran Associates Inc., Red Hook, NY, USA, NIPS'20

Fort G, Gach P, Moulines E (2021a) Fast Incremental Expectation Maximization for finite-sum optimization: nonasymptotic convergence. Stat Comput 31(4):48. https://doi.org/10.1007/s11222-021-10023-9

Fort G, Moulines E, Wai HT (2021b) Geom-Spider-EM: Faster Variance Reduced Stochastic Expectation Maximization for Nonconvex Finite-Sum Optimization. In: ICASSP 2021 - 2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp 3135–3139, https://doi.org/10.1109/ICASSP39728.2021.9414271

Ghadimi S, Lan G (2013) Stochastic First- and Zeroth-Order Methods for Nonconvex Stochastic Programming. SIAM Journal on Optimization 23(4):2341–2368. https://doi.org/10.1137/120880811

Ghadimi S, Lan G, Zhang H (2016) Mini-batch stochastic approximation methods for nonconvex stochastic composite optimization. Math Program 155(1-2):267–305. https://doi.org/10.1007/s10107-014-0846-1

Gower R, Goldfarb D, Richtarik P (2016) Stochastic Block BFGS: Squeezing More Curvature out of Data. In: Balcan MF, Weinberger KQ (eds) Proceedings of The 33rd International Conference on Machine Learning, Proceedings of Machine Learning Research, vol 48. PMLR, New York, New York, USA, pp 1869–1878

Hiriart-Urruty JB, Lemaréchal C (1996) Convex Analysis and Minimization Algorithms. Springer Verlag, Heidelberg, https://doi.org/https://doi.org/10.1007/978-3-662-02796-7, two volumes - 2nd printing

Horváth S, Lei L, Richtárik P, et al (2022) Adaptivity of Stochastic Gradient Methods for Nonconvex Optimization. SIAM Journal on Mathematics of Data Science 4(2):634–648. https://doi.org/10.1137/21M1394308

Johnson R, Zhang T (2013) Accelerating stochastic gradient descent using predictive variance reduction. In: Burges C, Bottou L, Welling M, et al (eds) Advances in Neural Information Processing Systems, vol 26. Curran Associates, Inc.

Karimi B, Wai HT, Moulines E, et al (2019) On the Global Convergence of (Fast) Incremental Expectation Maximization Methods. In: Wallach H, Larochelle H, Beygelzimer A, et al (eds) Advances in Neural Information Processing Systems 32. Curran Associates, Inc., p 2837–2847

Karimi H, Nutini J, Schmidt M (2016) Linear Convergence of Gradient and Proximal-Gradient Methods Under the Polyak-Łojasiewicz Condition. In: Frasconi P, Landwehr N, Manco G, et al (eds) Machine Learning and Knowledge Discovery in

Databases. Springer International Publishing, pp 795–811

Kolte R, Erdogdu M, Ozgur A (2015) Accelerating svrg via second-order information. In: Advances in Neural Information Processing Systems - Workshop OptML, pp 1–5

Lan G (2020) First-order and Stochastic Optimization Methods for Machine Learning. Springer Series in the Data Sciences, Springer International Publishing, https://doi.org/https://doi.org/10.1007/978-3-030-39568-1

Lee JD, Sun Y, Saunders MA (2014) Proximal Newton-Type Methods for Minimizing Composite Functions. SIAM Journal on Optimization 24(3):1420–1443. https://doi.org/10.1137/130921428

Li Z, Li J (2018) A Simple Proximal Stochastic Gradient Method for Nonsmooth Nonconvex Optimization. In: Bengio S, Wallach H, Larochelle H, et al (eds) Advances in Neural Information Processing Systems, vol 31. Curran Associates, Inc.

Li Z, Bao H, Zhang X, et al (2021) PAGE: A Simple and Optimal Probabilistic Gradient Estimator for Nonconvex Optimization. In: Meila M, Zhang T (eds) Proceedings of the 38th International Conference on Machine Learning, Proceedings of Machine Learning Research, vol 139. PMLR, pp 6286–6295

McLachlan G, Krishnan T (2008) The EM algorithm and extensions, 2nd edn. Wiley series in probability and statistics, Wiley, https://doi.org/10.1002/9780470191613

Metel M, Takeda A (2021) Stochastic proximal methods for non-smooth non-convex constrained sparse optimization. Journal of Machine Learning Research 22(115):1–36

Meyn S, Tweedie R (1993) Markov Chains and Stochastic Stability. Springer-Verlag, London, https://doi.org/https://doi.org/10.1007/978-1-4471-3267-7

Moreau J (1965) Proximité et dualité dans un espace hilbertien. Bulletin de la Société Mathématique de France 93:273–299. https://doi.org/10.24033/bsmf.1625

Moritz P, Nishihara R, Jordan M (2016) A Linearly-Convergent Stochastic L-BFGS Algorithm. In: Gretton A, Robert CC (eds) Proceedings of the 19th International Conference on Artificial Intelligence and Statistics, Proceedings of Machine Learning Research, vol 51. PMLR, pp 249–258

Neal RM, Hinton GE (1998) A View of the EM Algorithm that Justifies Incremental, Sparse, and other Variants. In: Jordan MI (ed) Learning in Graphical Models. Springer Netherlands, Dordrecht, p 355–368, https://doi.org/10.1007/978-94-011-5014-9_12

Ng SK, McLachlan GJ (2003) On the choice of the number of blocks with the incremental EM algorithm for the fitting of normal mixtures. Stat Comput 13(1):45–55. https://doi.org/10.1023/A:1021987710829

Nguyen L, Liu J, Scheinberg K, et al (2017) SARAH: A novel method for machine learning problems using stochastic recursive gradient. In: Precup D, Teh YW (eds) Proceedings of the 34th International Conference on Machine Learning, pp 2613–2621

Nhan HP, Lam MN, Dzung TP, et al (2020) ProxSARAH: An Efficient Algorithmic Framework for Stochastic Composite Nonconvex Optimization. Journal of Machine Learning Research 21(110):1–48

Park Y, Dhar S, Boyd S, et al (2019) Variable Metric Proximal Gradient Method with Diagonal Barzilai-Borwein Stepsize. https://doi.org/10.48550/ARXIV.1910.07056

Polson NG, Scott J, Windle J (2013) Bayesian Inference for Logistic Models Using P'olya–Gamma Latent Variables. Journal of the American Statistical Association 108(504):1339–1349. https://doi.org/https://doi.org/10.1080/01621459.2013.829001

Reddi SJ, Hefny A, Sra S, et al (2016) Stochastic Variance Reduction for Nonconvex Optimization. In: Balcan MF, Weinberger KQ (eds) Proceedings of The 33rd International Conference on Machine Learning, Proceedings of Machine Learning Research, vol 48. PMLR, New York, New York, USA, pp 314–323

Repetti A, Wiaux Y (2021) Variable metric forward-backward algorithm for composite minimization problems. SIAM J Optim 31:1215–1241. https://doi.org/https://doi.org/10.1137/19M1277552

Repetti A, Chouzenoux E, Pesquet JC (2014) A preconditioned forward-backward approach with application to large-scale nonconvex spectral unmixing problems. In: 2014 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp 1498–1502, https://doi.org/10.1109/ICASSP.2014.6853847

Robert C, Casella G (2004) Monte Carlo statistical methods. Springer Verlag, https://doi.org/https://doi.org/10.1007/978-1-4757-4145-2

Wang Z, Ji K, Zhou Y, et al (2019) Spiderboost and momentum: Faster variance reduction algorithms. In: Wallach HM, Larochelle H, Beygelzimer A, et al (eds) Advances in Neural Information Processing Systems 32: Annual Conference on Neural Information Processing Systems 2019, NeurIPS 2019, December 8-14, 2019, Vancouver, BC, Canada, pp 2403–2413

Wei G, Tanner M (1990) A Monte Carlo Implementation of the EM Algorithm and the Poor Man's Data Augmentation Algorithms. J Am Stat Assoc 85(411):699–704. https://doi.org/10.1080/01621459.1990.10474930

Wu C (1983) On the Convergence Properties of the EM Algorithm. Ann Statist 11(1):95–103. https://doi.org/10.1214/aos/1176346060

Yun J, Lozano AC, Yang E (2021) Adaptive proximal gradient methods for structured neural networks. In: Ranzato M, Beygelzimer A, Dauphin Y, et al (eds) Advances in Neural Information Processing Systems, vol 34. Curran Associates, Inc., pp 24,365–24,378

Zhang J, Xiao L (2019) A stochastic composite gradient method with incremental variance reduction. In: Wallach H, Larochelle H, Beygelzimer A, et al (eds) Advances in Neural Information Processing Systems, vol 32. Curran Associates, Inc.

Zhang Q, Huang F, Deng C, et al (2022) Faster stochastic quasi-newton methods. IEEE Transactions on Neural Networks and Learning Systems 33(9):4388–4397. https://doi.org/10.1109/TNNLS.2021.3056947

Zhou D, Xu P, Gu Q (2020) Stochastic Nested Variance Reduction for Nonconvex Optimization. Journal of Machine Learning Research 21(103):1–63