Noname manuscript No.
(will be inserted by the editor)

SABRINA: A Stochastic Subspace Majorization-Minimization Algorithm

Emilie Chouzenoux · Jean-Baptiste Fest

Received: date / Accepted: date

Abstract A wide class of problems involves the minimization of a coercive and differentiable function F on $\mathbb{R}^N$ whose gradient cannot be evaluated in an exact manner. In such context, many existing convergence results from standard gradient-based optimization literature cannot be directly applied and robustness to errors in the gradient is not necessarily guaranteed. This work is dedicated to investigating the convergence of Majorization-Minimization (MM) schemes when stochastic errors affect the gradient terms. We introduce a general stochastic optimization framework, called SABRINA (StochAstic suBspace majoRIzation-miNimization Algorithm) that encompasses MM quadratic schemes possibly enhanced with a subspace acceleration strategy. New asymptotical results are built for the stochastic process generated by SABRINA. Two sets of numerical experiments in the field of machine learning and image processing are presented to support our theoretical results and illustrate the good performance of SABRINA with respect to state-of-the-art gradient-based stochastic optimization methods.

Keywords Stochastic optimization, convergence analysis, Majorization-Minimization, subspace acceleration, binary logistic regression, image reconstruction.

.

1 Introduction

We consider the problem:

$$\underset{\boldsymbol{x}\in\mathbb{R}^N}{\text{minimize}} \quad F(\boldsymbol{x}), \tag{1}$$

where $F : \mathbb{R}^N \to \mathbb{R}$ is a coercive and differentiable function on $\mathbb{R}^N$. We focus on the case when the gradient of F is altered by stochastic errors during the iterative

E. Chouzenoux · J.-B. Fest (corresponding author)
Université Paris-Saclay, Inria, CentraleSupélec, Centre de Vision Numérique,
9 rue Joliot Curie
91190 Gif sur Yvette, France
Tel.:+33175316994
E-mail: jean-baptiste.fest@centralesupelec.fr

optimization process. This problem has been widely studied in the optimization literature, starting from seminal works [33, 61], and has known a renewed interest in the last decade with applicative challenges arising in supervised learning on large scale datasets [12, 54]. The stability properties of gradient-based stochastic schemes are also of high interest in approximate Bayesian inference, where stochastic gradient steps are often used to improve the exploration capacities of the samplers [32, 47, 51, 58].

Probably the most relevant gradient-based stochastic optimizer is the stochastic gradient descent (SGD) algorithm, studied in [7, 33, 61]. Extension of SGD to non-differentiable case using proximal-based tools can be found in [2, 4, 26, 46, 64]. Few convergence studies made in the deterministic case extend straightforwardly to the stochastic case. All the aforementioned works are grounded on specific probabilistic tools such as [31, 62]. SGD is rather simple but can exhibit slow convergence. Therefore, many recent works have focused on deriving accelerated variants of it. Two main families of acceleration strategies can be distinguished in the literature. The first approach, adopted for example in [30, 39, 45, 49, 55], relies on subspace (i.e., momentum) acceleration. The convergence rate is improved by using information from past iterates for the construction of new estimates. The second approach to accelerate the convergence of SGD is based on a variable metric strategy [19, 29]. The underlying metric is modified at each iteration thanks to a preconditioning matrix, which may incorporate second-order information about the function to minimize. These acceleration techniques give rise to promising practical results.

This work proposes a novel SGD-based scheme to solve Problem (1), by combining the two aforementioned acceleration strategies. To do so, we rely on the so-called Majorization-Minimization (MM) principle [69, 73]. At each iteration of an MM algorithm, a surrogate function majorizing the problem cost function is constructed. The next iterate is then obtained by minimizing the majorant. By construction, MM method produces a sequence of iterates that decreases the cost function monotonically. MM algorithms benefit from assessed convergence properties in the convex and non-convex settings [9, 23, 44]. The extension of MM methodology to the stochastic context has been studied recently in [25, 27, 50] in restricted scenarios. The method proposed by [27] is dedicated to introducing stochastic errors into the expectation-minimization approach, a special case of MM. The MISO approach from [50] combines an MM scheme with constraining averaging rules both over surrogates and iterates to reach convergence. The work of [25] studies a scheme close to the one proposed in our paper, but limits the analysis to the specific case of a penalized least-square criterion whose gradient is evaluated using a recursive least-squares implementation [35]. In this present work, we introduce a versatile MM scheme relying on quadratic majorant surrogates for F and allowing for subspace acceleration [22, 60]. In a nutshell, the resulting algorithm benefits from a simple structure that can be understood as an SGD method with both preconditioning and momentum-based term, and has minimal parameter tuning. For the proposed scheme, our contributions are[1]:

[1] A preliminary version of this work has been presented in the conference proceedings [36]. The convergence result was weaker, and stated without proof. The experimental validation was limited to a single, simpler, numerical scenario.

- almost sure convergence results for non necessarily convex F;
- convergence rate analysis in the strongly convex case;
- illustration of the performance and comparison with state-of-the-art on two numerical examples.

The rest of the article is organized as follows. Section 2 states notations and introduces the considered MM stochastic optimization scheme. The probabilistic framework is introduced in Section 3.1. Assumptions are listed in Section 3.2 along with a discussion in Section 3.3. Some technical lemmas, essential for our theoretical study, are presented in Sections 3.4 and 3.5. Our main contribution is concentrated in Section 4. Our main convergence results can be found in Theorem 1 (Section 4.2) and Theorem 2 (Section 4.3). Numerical experiments are provided in Sections 5 and 6. Finally, we conclude the paper in Section 7.

2 Background and proposed formulation

2.1 Notations

We classically denote by $\|\cdot\|^2 = \langle\cdot|\cdot\rangle$ the euclidean norm of $\mathbb{R}^N$, and $|||\cdot|||$ the spectral norm (i.e., largest singular value) of elements of $\mathbb{R}^{M\times N}$. If $\boldsymbol{M}$ is a symmetric definite positive matrix of $\mathbb{R}^{N\times N}$, $\|.\|^2_{\boldsymbol{M}}$ corresponds to $\langle.|\boldsymbol{M}.\rangle$. Moreover, we will use the Loewner order $\preceq$ [8, Ch.V] between two symmetric matrices $\boldsymbol{M}_1$, $\boldsymbol{M}_2$ of $\mathbb{R}^{N\times N}$, where relation $\boldsymbol{M}_1 \preceq \boldsymbol{M}_2$ holds if and only if difference $\boldsymbol{M}_2 - \boldsymbol{M}_1$ is (symmetric) positive. $\boldsymbol{I}_N$ states for the identity matrix of $\mathbb{R}^N$, $\boldsymbol{0}_N$ the zero vector of size N, and $\boldsymbol{O}_N$ the null matrix of $\mathbb{R}^{N\times N}$. Bold symbols are used for matrix and vectors. Italic style is retained for deterministic quantities. Ker and ran denote the kernel and range (i.e., image) respectively of a linear operator.

Subject to existence, $\widetilde{\boldsymbol{x}}$ will state for a stationary point of F. Moreover, $\operatorname{zer}\nabla F$ will denote the set of stationary points of F. We will write $\boldsymbol{x}^*$ a global minimizer for F and define $F^* := F(\boldsymbol{x}^*)$.

2.2 Quadratic MM approach

MM algorithm solves Problem (1) iteratively by generating a sequence $(\boldsymbol{x}_k)_{k\in\mathbb{N}}$ of elements of $\mathbb{R}^N$, where the step from the iterate $\boldsymbol{x}_k$ to its successor $\boldsymbol{x}_{k+1}$ is achieved through the minimization of $h(\cdot, \boldsymbol{x}_k)$, a tangent majorant surrogate of F around $\boldsymbol{x}_k$, i.e.

$$(\forall \boldsymbol{x} \in \mathbb{R}^N) \quad h(\boldsymbol{x}, \boldsymbol{x}_k) \geq F(\boldsymbol{x}) \quad \text{and} \quad h(\boldsymbol{x}_k, \boldsymbol{x}_k) = F(\boldsymbol{x}_k). \tag{2}$$

An efficient strategy consists in resorting to a quadratic majorant function, structurally analogous to a second-order Taylor's expansion of F:

$$h : (\boldsymbol{x}, \boldsymbol{y}) \mapsto F(\boldsymbol{y}) + \nabla F(\boldsymbol{y})^\top (\boldsymbol{x} - \boldsymbol{y}) + \frac{1}{2}\|\boldsymbol{x} - \boldsymbol{y}\|^2_{\boldsymbol{A}(\boldsymbol{y})}. \tag{3}$$

Hereabove, for every $\boldsymbol{y} \in \mathbb{R}^N$, $\boldsymbol{A}(\boldsymbol{y})$ is a symmetric positive definite matrix of $\mathbb{R}^{N\times N}$ chosen so as to ensure (2). The latter, called the majorant metric matrix,

yields a complete description of $h(\cdot, \boldsymbol{y})$ and thus influences the approximation quality of F by this same surrogate. Several techniques for building suitable majorant metric matrices can be found for a wide class of problems encompassing image restoration, telecommunication or supervised learning in [23, 69, 73].

As a consequence of the invertibility of $\boldsymbol{A}(\boldsymbol{x}_k)$, for every $k \in \mathbb{N}$, we obtain the generic MM scheme [69]:

$$\begin{aligned}(\forall k \in \mathbb{N}) \quad \boldsymbol{x}_{k+1} &= \operatorname{argmin}_{\boldsymbol{x}\in\mathbb{R}^N} \quad h(\boldsymbol{x}, \boldsymbol{x}_k), \\ &= \boldsymbol{x}_k - \boldsymbol{A}_k^{-1}\nabla F(\boldsymbol{x}_k) \end{aligned} \tag{4}$$

with $\boldsymbol{A}_k := \boldsymbol{A}(\boldsymbol{x}_k)$ and $\boldsymbol{x}_0 \in \mathbb{R}^N$. The MM update (4) can be shown to map with the half-quadratic algorithm [40] when F is a penalized least-squares function. By construction, the sequence $(\boldsymbol{x}_k)_{k\in\mathbb{N}}$ built by (4) guarantees a monotonic decrease of $(F(\boldsymbol{x}_k))_{k\in\mathbb{N}}$. Convergence of $(\boldsymbol{x}_k)_{k\in\mathbb{N}}$ to a stationary point of F can be shown under suitable technical assumptions on F and $(\boldsymbol{A}_k)_{k\in\mathbb{N}}$ [3].

2.3 Subspace acceleration

When using update (4), one needs to invert an $N \times N$ matrix. Such an operation is undesirable when N is large. The authors from [22] proposed to integrate a so-called subspace acceleration procedure [60, 72] into (4) leading to:

$$(\forall k \in \mathbb{N}) \qquad \boldsymbol{x}_{k+1} = \boldsymbol{x}_k + \boldsymbol{D}_k \boldsymbol{u}_k, \tag{5}$$

with

$$(\forall k \in \mathbb{N}) \quad \boldsymbol{u}_k \in \underset{\boldsymbol{u}\in\mathbb{R}^{M_k}}{\arg\min}\, h(\boldsymbol{x}_k + \boldsymbol{D}_k\boldsymbol{u}, \boldsymbol{x}_k), \tag{6}$$

and $\boldsymbol{x}_0 \in \mathbb{R}^N$. The key ingredient of the above method is the introduction of a matrix $\boldsymbol{D}_k \in \mathbb{R}^{N\times M_k}$ with $N \geq M_k \geq 1$, which imposes a subspace to search for the new iterate $\boldsymbol{x}_{k+1}$. Taking $M_k = N$ and $\boldsymbol{D}_k = \boldsymbol{I}_N$, the identity matrix of $\mathbb{R}^N$, (5) goes back to scheme (4). In practice, only a few degrees of freedom are actually required to reach good convergence speed (see [24] for a detailed analysis of the convergence rate of scheme (5) as a function of $\boldsymbol{D}_k$ and $\boldsymbol{A}_k$), so M_k is typically retained as very small compared to N. Interesting choices can be found in [22, Tab.1]. Setting $\boldsymbol{D}_k = [-\nabla F(\boldsymbol{x}_k) \quad | \quad \boldsymbol{x}_k - \boldsymbol{x}_{k-1}]$ (with convention $\boldsymbol{x}_{-1} = \boldsymbol{x}_0$) brings notably to the so-called MM Memory Gradient (3MG) method whose great performances have been illustrated in [22, 23, 37, 66]. Other choices for the subpace matrix can be found in [53, 60, 67, 74]. It is worth noting that the minimization scheme (5) shares strong connections with non-linear conjugate gradient algorithm [56], low-memory quasi-Newton approaches such as L-BFGS [48, 56], trust-region strategies [1], and momentum-based methods [70]. In contrast with these aforementioned works, the MM subspace scheme presents the key advantage of a simple linesearch procedure (6) associated with sounded convergence guarantees. Indeed, assuming, without loss of generality that $\boldsymbol{D}_k$ has full column rank, the quadratic structure of $h(\cdot, \boldsymbol{x}_k)$ allows to obtain an analytical solution to sub-problem (6).

$$(\forall k \in \mathbb{N}) \quad \boldsymbol{u}_k = -\left(\boldsymbol{D}_k^\top \boldsymbol{A}_k \boldsymbol{D}_k\right)^{-1} \boldsymbol{D}_k^\top \nabla F(\boldsymbol{x}_k). \tag{7}$$

The interest lies here in the fact that $\boldsymbol{D}_k^\top \boldsymbol{A}_k \boldsymbol{D}_k$ is an $M_k \times M_k$ matrix making its inversion far easier computable than the inversion of $\boldsymbol{A}_k$, as soon as M_k is small. Convergence properties of (5)-(6) have been established in the convex setting in [22], and extended to the non-convex setting in [23] using recent tools of non smooth analysis. The catalizing effect of the subspace acceleration for practical convergence speed of MM methods has been acknowledged in the survey paper [69]. We also refer the reader to [17,20,41] for practical implementation of MM subspace approaches on modern high performance computing tools.

2.4 SABRINA, a stochastic subspace MM algorithm

We are now ready to introduce the algorithm studied in this paper. We focus on the stability of the optimization scheme (5)-(6) when the gradient of F is affected by an additive stochastic perturbation at each iteration $k \in \mathbb{N}$, so that only the approximate value $\mathbf{g}_k$, defined below, is available:

$$(\forall k \in \mathbb{N}) \quad \mathbf{g}_k = \nabla F(\mathbf{x}_k) + \mathbf{e}_k. \tag{8}$$

Hereabove, $(\mathbf{e}_k)_{k\in\mathbb{N}}$ corresponds to a zero-mean stochastic process with a bounded variance in a sense that will be specified in Section 3.2. Formulating the stochastic counterpart of (5)-(6) requires to introduce the concept of inexact majorant function. For every $k \in \mathbb{N}$, the majorant function $h(\cdot, \mathbf{x}_k)$ will be substituted by a new function $\widehat{\mathrm{h}}_k$ with the following expression:

$$\widehat{\mathrm{h}}_k : \boldsymbol{u} \in \mathbb{R}^N \mapsto F(\mathbf{x}_k) + \mathbf{g}_k^\top (\boldsymbol{u} - \mathbf{x}_k) + \frac{1}{2}\|\boldsymbol{u} - \mathbf{x}_k\|^2_{\mathbf{A}_k}. \tag{9}$$

In analogy with the deterministic formulation from Section 2.3, the update at iteration $k \in \mathbb{N}$ will be grounded on the search of a minimizer of $\widehat{\mathrm{h}}_k$ along the directions spanned by the columns of a matrix $\mathbf{D}_k \in \mathbb{R}^{M_k \times N}$.

Let us also introduce a positive stepsize sequence $(\gamma_k)_{k\in\mathbb{N}}$ in order to promote stability of the iterates. This finally leads us to our stochastic minimization scheme called SABRINA (StochAstic suBspace majoRization mINimization Algorithm):

$$(\forall k \in \mathbb{N}) \quad \mathbf{x}_{k+1} = \mathbf{x}_k + \gamma_k \mathbf{D}_k \mathbf{u}_k, \tag{10}$$

with

$$(\forall k \in \mathbb{N}) \quad \mathbf{u}_k = -\left(\mathbf{D}_k^\top \mathbf{A}_k \mathbf{D}_k\right)^{-1} \mathbf{D}_k^\top \mathbf{g}_k, \tag{11}$$

and $\mathbf{x}_0 \equiv \boldsymbol{x}_0 \in \mathbb{R}^N$, a deterministic quantity.

Remark 1 For the sake of clarity, throughout the paper, we distinguish deterministic and random quantities, with italic and non-italic styles, respectively. In particular, since the noise $(\mathbf{e}_k)_{k\in\mathbb{N}}$ is random, the quantities $(\mathbf{g}_k, \mathbf{x}_k, \mathbf{D}_k, \mathbf{u}_k)_{k\in\mathbb{N}}$ are too. The probabilistic notations (i.e., probability space, filtration), useful for our theoretical analysis, will be made explicit in Sec. 3.1.

2.5 Link with stochastic preconditioned gradient algorithm

It is straightforward to rewrite SABRINA iterations (10)-(11) under the compact form:

$$(\forall k \in \mathbb{N}) \quad \mathbf{x}_{k+1} = \mathbf{x}_k - \gamma_k \mathbf{B}_k \mathbf{g}_k, \tag{12}$$

with

$$(\forall k \in \mathbb{N}) \quad \mathbf{B}_k = \mathbf{D}_k \left(\mathbf{D}_k^\top \mathbf{A}_k \mathbf{D}_k \right)^{-1} \mathbf{D}_k^\top. \tag{13}$$

The above formulation is interesting as it highlights similarities between SABRINA and the preconditioned gradient scheme with inexact gradient term, studied for instance in [11, 16]. The main distinction is that the symmetric matrix $\mathbf{B}_k \in \mathbb{R}^{N \times N}$ involved in (12) gathers information brought by the majorant matrix $\mathbf{A}_k$ and by the retained subspace $\mathbf{D}_k$, as described in (13). The formulation above suggests that controlling the behaviour of $(\mathbf{x}_k)_{k\in\mathbb{N}}$ requires studying the properties of $(\mathbf{B}_k)_{k\in\mathbb{N}}$, which raises two main theoretical challenges that we plan to tackle in this work: (i) $\mathbf{B}_k$ is a random matrix with rank lower or equal than M_k, (ii) F is not assumed to be a convex function. Up to our knowledge, the general scheme (12) has never been analysed under these two restrictions.

3 Preliminary lemmas

In this section, we introduce our probabilistic notations. We present and discuss our assumptions. Finally, we prove three technical lemmas that appear essential for establishing our main convergence results presented in Section 4.

3.1 Probabilistic framework

In the remainder of the paper, we consider $(\Omega, \mathcal{F}, P)$ a probability space to which we associate the filtration $(\mathcal{F}_k)_{k\in\mathbb{N}}$ where $\mathcal{F}_0 = \{\Omega, \emptyset\}$ and for all $k \geq 1$, $\mathcal{F}_k = \sigma\left(\mathbf{e}_0, \mathbf{x}_1, ..., \mathbf{e}_{k-1}, \mathbf{x}_k\right)$ corresponds to the sub-sigma algebra generated by the family $\{\mathbf{e}_0, \mathbf{x}_1, ..., \mathbf{e}_{k-1}, \mathbf{x}_k\}$ of random variables. For each $k \in \mathbb{N}$, $\mathcal{F}_k$ gathers all the information available from the origin of the process to iteration k. A mathematical property will be said to be verified *almost surely* or *a.s.* if it holds on a probability-one set belonging to $\mathcal{F}$. We also remind that an element of $\mathcal{F}$ is negligible if it is a probability-zero one. For a given $k \in \mathbb{N}$ and subject to existence, we will denote $\mathbb{E}(.|\mathcal{F}_k)$, the conditional expectancy operator associated to $\mathcal{F}_k$.

3.2 Assumptions

The following assumptions will guide us throughout the rest of the study.

Assumption 1 *F is coercive and β-Lipschitz differentiable on $\mathbb{R}^N$, i.e. there exists $\beta > 0$ such that:*

$$(\forall (\boldsymbol{x}, \boldsymbol{y}) \in (\mathbb{R}^N)^2) \quad \|\nabla F(\boldsymbol{x}) - \nabla F(\boldsymbol{y})| \leq \beta \, \|\boldsymbol{x} - \boldsymbol{y}\|. \tag{14}$$

Assumption 2 *There exists* $(\eta, \nu) > 0$ *such that:*

$$(\forall k \in \mathbb{N}) \quad \eta \boldsymbol{I}_N \preceq \mathbf{A}_k \preceq \nu \boldsymbol{I}_N \quad a.s. \tag{15}$$

Assumption 3 *For every iteration* $k \in \mathbb{N}$*,*

(i) $\operatorname{rank}(\mathbf{D}_k) = M_k \quad a.s.$
(ii) $\mathbf{g}_k \in \operatorname{ran}(\mathbf{D}_k) \quad a.s.$

Assumption 4 *The stochastic noise process* $(\mathbf{e}_k)_{k\in\mathbb{N}}$ *fulfills:*

(i) $(\forall k \in \mathbb{N}) \;\; \mathbb{E}\left(\mathbf{e}_k | \mathcal{F}_k\right) = 0 \quad a.s.$
(ii) There exists $C \in (0, C_{\max})$ *with* $C_{\max} = \frac{1}{2}\left((1 + \frac{4\eta}{\nu})^{\frac{1}{2}} - 1\right)$ *such that:*

$$(\forall k \in \mathbb{N}) \;\; \mathbb{E}\left(\|\mathbf{e}_k\|^2 | \mathcal{F}_k\right) \leq C^2 \|\nabla F(\mathbf{x}_k)\|^2 \quad a.s. \tag{16}$$

Assumption 5 $(\gamma_k)_{k\in\mathbb{N}}$ *is a sequence of strictly positive scalars satisfying:*

$$\gamma_k \underset{k\to+\infty}{\longrightarrow} 0 \;\; \textit{and} \;\; \sum_{k=0}^{+\infty} \gamma_k = +\infty.$$

3.3 Discussion on the assumptions

Assumption 1 is rather standard in the analysis of stochastic gradient-based methods [42,50]. It is worth noting that the knowledge of the Lipschitz constant of ∇F is not necessary for the practical implementation of the method.

Assumption 2 is essential for ensuring convergence of MM methods involving quadratic majorant functions, as it ensures that the majorant metric matrices remain well-conditioned. Let us remark that the existence of such matrices is guaranteed by the descent lemma, since one can set $\mathbf{A}_k \equiv \beta \boldsymbol{I}_N$, with β the Lipschitz constant of ∇F (see Assumption 1). For such choice, SABRINA identifies with SGD with specific MM-based closed-form formulas for the stepsize and the momentum weight. As we will show in our experimental tests, it is however usually worthy to search for more sophisticated choices for $(\mathbf{A}_k)_{k\in\mathbb{N}}$, leading usually to faster practical convergence (See also [3, Sec.IV], [24] for the role of majorant mappings in the convergence speed of quadratic MM methods).

Assumptions 3(i) and 3(ii) work as a peer, and control the validity of the subspace construction. These requirements are standard in subspace-based optimization methods [22,60,72]. Assumption 3(i) ensures the non-redundancy of the information within the subspace. Assumption 3(ii) enhances some descent properties of the algorithm. Note that the latter Assumption is verified as soon as one of the columns of $\mathbf{D}_k$ identifies with $-\mathbf{g}_k$ (i.e. the SGD direction). A practical way to build $\mathbf{D}_k$ satisfying Assumption 3(ii) is thus to set its first column as $-\mathbf{g}_k$, and to add extra columns including, for example, difference on past iterates/directions. A list of valid subspace constructions can be found in [25, Tab.II]. A (column) rank reduction is performed, so as to satisfy Assumption 3(i). Interestingly, for $\mathbf{D}_k \equiv -\mathbf{g}_k$, SABRINA reads as a preconditioned SGD algorithm, with MM-based preconditioner.

Assumption 4(i) is often required for studying the stability of gradient-based optimization schemes in the presence of stochastic errors [28,42]. Assumption 4(ii) corresponds to a second order moment property and can be seen as a particular case of [13, Assumption 4.3.c]. It states that uncertainty $\mathbf{e}_k$ should remain reasonable with respect to the norm of the (true) gradient of F at $\mathbf{x}_k$. The larger condition number η/ν of the majorant metrics, the more permissive upper bound $C_{\max}$ is. The maximum theoretical bound $\frac{\sqrt{5}-1}{2} \simeq 6.18 \times 10^{-1}$ is reached if and only if $\eta \equiv \nu$. Such a situation occurs for instance when $\mathbf{A}_k$ equals to a positive constant times identity. Typical choice would be $\mathbf{A}_k \equiv \beta \boldsymbol{I}_N$, but, as already mentioned, this choice might be detrimental to the convergence speed. In contrast, one can easily show that $C_{\max} \sim \eta/\nu$ for $\eta/\nu \to 0^+$, which means that poorly conditioned majorant mappings would demand a high level of precision on the gradient's uncertainty. This suggests that a compromise must be achieved between the convergence speed and the requirements in terms of stability to noise.

Assumption 5 is a relaxed version of the classical σ-sequence hypothesis [38]. In particular, a main feature of our study is that it is not necessary to impose the usual condition $\sum_{k=0}^{+\infty} \gamma_k^2 < +\infty$. Assumption 5 allows to choose a stepsize $(\gamma_k)_{k\in\mathbb{N}}$ with a slow convergence to 0 (e.g., an inverse logarithmic one).

3.4 Properties of the preconditioning matrices

As mentioned in Section 2.5, the behaviour of SABRINA iterates depends on the properties of $(\mathbf{B}_k)_{k\in\mathbb{N}}$ expressed in (13). We derive some useful technical properties for these matrices, gathered in the lemma below.

Lemma 1 *Under Assumptions 2 and 3(i), for all $k \in \mathbb{N}$, $\mathbf{B}_k$ is almost surely well-defined and satisfies:*

$$\mathbf{D}_k \mathbf{u}_k = -\mathbf{B}_k \mathbf{g}_k, \tag{17a}$$

$$\boldsymbol{O}_N \preceq \mathbf{B}_k \preceq \frac{1}{\eta} \boldsymbol{I}_N, \tag{17b}$$

$$(\forall \boldsymbol{x} \in \mathrm{ran}(\mathbf{D}_k)) \quad \boldsymbol{x}^\top \mathbf{B}_k \boldsymbol{x} \geq \frac{1}{\nu} \|\boldsymbol{x}\|^2. \tag{17c}$$

Proof Let $k \in \mathbb{N}$.

Matrix $\mathbf{D}_k^\top \mathbf{A}_k \mathbf{D}_k$ is symmetric. Using Loewner order properties [8, Ch. V] and Assumption 2, we almost surely have [2] :

$$\eta \mathbf{D}_k^\top \mathbf{D}_k \preceq \mathbf{D}_k^\top \mathbf{A}_k \mathbf{D}_k \preceq \nu \mathbf{D}_k^\top \mathbf{D}_k. \tag{18}$$

Assumption 3(i) ensures that $\mathbf{D}_k$ is an injective operator. It follows that $\mathbf{D}_k^\top \mathbf{D}_k$ is a symmetric definite positive matrix and according to (18) and $\eta > 0$, so is $\mathbf{D}_k^\top \mathbf{A}_k \mathbf{D}_k$. This ensures that $\mathbf{B}_k$, as defined in (13), exists. Then, (17a) directly comes from (10) and (12).

[2] If $A \preceq B$ and D is a non necessary square matrix, then $D^\top A D \preceq D^\top B D$.

Moreover, since the three terms in (18) are invertible matrices, we have:

$$\frac{1}{\nu}(\mathbf{D}_k^\top \mathbf{D}_k)^{-1} \preceq (\mathbf{D}_k^\top \mathbf{A}_k \mathbf{D}_k)^{-1} \preceq \frac{1}{\eta}(\mathbf{D}_k^\top \mathbf{D}_k)^{-1}, \tag{19}$$

so that (by footnote 2):

$$\frac{1}{\nu}\mathbf{D}_k(\mathbf{D}_k^\top \mathbf{D}_k)^{-1}\mathbf{D}_k^\top \preceq \mathbf{B}_k \preceq \frac{1}{\eta}\mathbf{D}_k(\mathbf{D}_k^\top \mathbf{D}_k)^{-1}\mathbf{D}_k^\top. \tag{20}$$

Let us denote:

$$\mathbf{P}_k = \mathbf{D}_k(\mathbf{D}_k^\top \mathbf{D}_k)^{-1}\mathbf{D}_k^\top. \tag{21}$$

$\mathbf{P}_k \in \mathbb{R}^{N\times N}$ is an orthogonal projection operator since it is symmetric and verifies $\mathbf{P}_k^2 = \mathbf{P}_k$. This latter equality can be rewritten as $\mathbf{P}_k = \mathbf{P}_k^\top \boldsymbol{\Delta} \mathbf{P}_k$ with $\boldsymbol{\Delta}$ a diagonal matrix of $\mathbb{R}^{N\times N}$ with only binary entries, so that $\boldsymbol{O}_N \preceq \boldsymbol{\Delta} \preceq \boldsymbol{I}_N$. Using again footnote 2, it follows that:

$$\boldsymbol{O}_N \preceq \mathbf{P}_k \preceq \boldsymbol{I}_N. \tag{22}$$

(17b) is then directly obtained by replacing (22) in (20).

As an orthogonal projection matrix, $\mathbf{P}_k$ satisfies

$$(\forall \boldsymbol{x} \in \mathrm{Ker}(\mathbf{P}_k)^\perp) \quad \mathbf{P}_k \boldsymbol{x} = \boldsymbol{x}. \tag{23}$$

Combining (23) with the left inequality of (20) yields:

$$(\forall \boldsymbol{x} \in \mathrm{Ker}(\mathbf{P}_k)^\perp) \quad \boldsymbol{x}^\top \mathbf{B}_k \boldsymbol{x} \geq \frac{1}{\nu}\boldsymbol{x}^\top \mathbf{P}_k \boldsymbol{x} = \frac{1}{\nu}\|\boldsymbol{x}\|^2. \tag{24}$$

There remains to show the relation $\mathrm{ran}(\mathbf{D}_k) = \mathrm{Ker}(\mathbf{P}_k)^\perp$. To do so, we rely on the classical linear algebra relation $\mathrm{ran}(\mathbf{D}_k)^\perp = \mathrm{Ker}(\mathbf{D}_k^\top)$ and thus prove $\mathrm{Ker}(\mathbf{D}_k^\top) = \mathrm{Ker}(\mathbf{P}_k)$ instead. Inclusion $\mathrm{Ker}(\mathbf{D}_k^\top) \subset \mathrm{Ker}(\mathbf{P}_k)$ is straightforward. Since $\boldsymbol{x} \in \mathrm{Ker}(\mathbf{P}_k)$, from the expression of $\mathbf{P}_k$ and left multiplication by $\boldsymbol{x}^\top$, we have

$$\boldsymbol{x}^\top \mathbf{D}_k(\mathbf{D}_k^\top \mathbf{D}_k)^{-1}\mathbf{D}_k^\top \boldsymbol{x} = 0. \tag{25}$$

Since $\mathbf{D}_k^\top \mathbf{D}_k$ is definite positive matrix, its inverse is too, so that $\mathbf{D}_k^\top \boldsymbol{x} = \mathbf{0}$, i.e. $\boldsymbol{x} \in \mathrm{Ker}(\mathbf{D}_k^\top)$ which concludes the proof of (17c).

□

Relation (17c) brings light into our interpretation of Assumption 3(ii) as a descent condition. Indeed, taking $\boldsymbol{x} = -\mathbf{g}_k$ in (17c) leads to the gradient-related inequality [6] considered for instance in the analysis of [22, 23]. Relation (17c) will actually play a key role in the asymptotical analysis of Section 4.

3.5 Two additional technical lemmas

The next lemma is essential as it guarantees the integrability of all the probabilistic quantities we will manipulate in our convergence analysis. It especially validates the use of the conditional expectation operator and of its associate properties in every situation encountered in our proofs.

Lemma 2 *Under Assumptions 1, 2, 3(i) and 4(ii), for every $k \in \mathbb{N}$, $\mathbf{x}_k$, $\nabla F(\mathbf{x}_k)$, $\mathbf{e}_k$ and $\mathbf{g}_k$ are square-integrable random vectors of $\mathbb{R}^N$. Moreover, $F(\mathbf{x}_k)$ is an integrable random variable of $\mathbb{R}$.*

Proof First, according to Assumption 1, F is a differentiable and coercive function on $\mathbb{R}^N$, which ensures the existence of a global minimizer $\boldsymbol{x}^*$ satisfying $\nabla F(\boldsymbol{x}^*) = \mathbf{0}_N$. Let us denote by F^* the minimal value of F on $\mathbb{R}^N$, i.e. $F^* = F(\boldsymbol{x}^*)$.

We start by proving the desired property for sequence $(\mathbf{x}_k)_{k\in\mathbb{N}}$. We here proceed by induction.

The case $k = 0$ is straightforward as $\mathbf{x}_0$ is a deterministic variable.

Assume that $\mathbf{x}_k$ is square-integrable for a given $k \in \mathbb{N}$. Then almost surely, and using Lemma 1,

$$\|\mathbf{x}_{k+1}\|^2 = \|\mathbf{x}_k - \gamma_k \mathbf{B}_k \mathbf{g}_k\|^2 \tag{26}$$
$$\leq 2\|\mathbf{x}_k\|^2 + 2\gamma_k^2 \|\mathbf{B}_k \mathbf{g}_k\|^2 \tag{27}$$
$$\leq 2\|\mathbf{x}_k\|^2 + 2\frac{\gamma_k^2}{\eta^2}\|\mathbf{g}_k\|^2. \tag{28}$$

with

$$\|\mathbf{g}_k\|^2 = \|\nabla F(\mathbf{x}_k) + \mathbf{e}_k\|^2 \tag{29}$$
$$\leq 2\|\nabla F(\mathbf{x}_k)\|^2 + 2\|\mathbf{e}_k\|^2. \tag{30}$$

Hereabove, the positivity of all the manipulated random variables makes possible to take the conditional expectations. Since $\nabla F(\mathbf{x}_k)$ is $\mathcal{F}_k$-measurable, the next inequalities follow by using Assumptions 1 and 3(i), almost surely

$$\mathbb{E}(\|\mathbf{g}_k\|^2|\mathcal{F}_k) = \mathbb{E}(\|\nabla F(\mathbf{x}_k) + \mathbf{e}_k\|^2|\mathcal{F}_k), \tag{31}$$
$$\leq 2\ \mathbb{E}(\|\nabla F(\mathbf{x}_k)\|^2|\mathcal{F}_k) + 2\ \mathbb{E}(\|\mathbf{e}_k\|^2|\mathcal{F}_k), \tag{32}$$
$$= 2\|\nabla F(\mathbf{x}_k)\|^2 + 2\ \mathbb{E}(\|\mathbf{e}_k\|^2|\mathcal{F}_k), \tag{33}$$
$$\leq 2(1 + C^2)\|\nabla F(\mathbf{x}_k)\|^2, \tag{34}$$
$$\leq 2\beta^2(1 + C^2)\|\mathbf{x}_k - \boldsymbol{x}^*\|^2, \tag{35}$$
$$\leq 4\beta^2(1 + C^2)(\|\mathbf{x}_k\|^2 + \|\boldsymbol{x}^*\|^2). \tag{36}$$

Taking the expectations yields

$$\mathbb{E}\left[\|\mathbf{g}_k\|^2\right] = \mathbb{E}\left[\mathbb{E}(\|\mathbf{g}_k\|^2|\mathcal{F}_k)\right] \tag{37}$$
$$\leq 4(1 + C^2)\beta^2(\mathbb{E}\left[\|\mathbf{x}_k\|^2\right] + \|\boldsymbol{x}^*\|^2). \tag{38}$$

By the induction hypothesis, we have $\mathbb{E}\left[\|\mathbf{x}_k\|^2\right] < +\infty$, so that using (28)-(38)

$$\mathbb{E}[\|\mathbf{x}_{k+1}\|^2] \leq 2\ \mathbb{E}\left[\|\mathbf{x}_k\|^2\right] + 8\beta^2\frac{\gamma_k^2}{\eta^2}(1+C^2)\left(\mathbb{E}\left[\|\mathbf{x}_k\|^2\right] + \|\boldsymbol{x}^*\|^2\right) \tag{39}$$

$$< +\infty, \tag{40}$$

which concludes this part of the proof.

We now focus on $\mathbf{g}_k$. The developments above shown that $\mathbb{E}\left[\|\mathbf{g}_k\|^2\right]$ is upper-bounded by a positive affine function of $\mathbb{E}\left[\|\mathbf{x}_k\|^2\right]$, itself being strictly lower than $+\infty$. Consequently, $\mathbb{E}\left[\|\mathbf{g}_k\|^2\right] < +\infty$.

Regarding $\nabla F(\mathbf{x}_k)$, we almost surely have

$$\|\nabla F(\mathbf{x}_k)\|^2 \leq \beta^2\|\mathbf{x}_k - \boldsymbol{x}^*\|^2 \tag{41}$$

$$\leq 2\beta^2(\|\mathbf{x}_k\|^2 + \|\boldsymbol{x}^*\|^2). \tag{42}$$

The right member in the above equation is integrable, and so is the same for $\|\nabla F(\mathbf{x}_k)\|^2$.

The integrability of $\|\mathbf{e}_k\|^2$ arises directly from Assumption 4(ii), passing directly to the expectation.

The descent lemma applied to F, which is a β-Lipschitz differentiable function of $\mathbb{R}^N$ according to Assumption 1, leads to

$$F(\mathbf{x}_k) - F^* \leq \frac{\beta}{2}\|\mathbf{x}_k - \boldsymbol{x}^*\|^2 \tag{43}$$

$$\leq \beta(\|\mathbf{x}_k\|^2 + \|\boldsymbol{x}^*\|^2). \tag{44}$$

The integrability of the right member of the above inequality yields the integrability of $F(\mathbf{x}_k)$.

□

We end this section with one last technical result which provides a rational for the expression of the bound $C_{\max}$ introduced in Assumption 4.

Lemma 3 *For every $C \in (0, C_{\max})$, there exists $\rho_0 > 0$ such that P_{ρ_0} is strictly negative on $[0, C]$ where for all $\rho > 0$, P_ρ refers to the polynomial*

$$P_\rho(X) = \left(1 + \frac{\nu\rho}{2\eta}\right)\frac{X^2}{\eta} + \frac{X}{\eta} + \left(\frac{\nu\rho}{2\eta^2} - \frac{1}{\nu}\right). \tag{45}$$

Proof For all $\rho > 0$, P_ρ is a second order polynomial whose discriminant Δ_ρ is

$$\Delta_\rho = \frac{1}{\eta^2} + \frac{4}{\eta}\left(1 + \frac{\nu\rho}{2\eta}\right)\left(\frac{1}{\nu} - \frac{\nu\rho}{2\eta^2}\right). \tag{46}$$

Taking $\rho \in \left(0, 2\,(\eta/\nu)^2\right)$, it follows that Δ_ρ is strictly positive. Thus, P_ρ admits two distinct roots

$$w_{\rho,1} = -\frac{\eta^2\sqrt{\Delta_\rho} + \eta}{\nu\rho + 2\eta} < 0, \quad \text{and} \quad w_{\rho,2} = \frac{\eta^2\sqrt{\Delta_\rho} - \eta}{\nu\rho + 2\eta}. \tag{47}$$

Taking the limit for vanishing ρ yields:

$$\lim_{\rho \to 0^+} w_{\rho,2} = \frac{\eta\sqrt{\frac{1}{\eta^2} + \frac{4}{\eta\nu}} - 1}{2} \tag{48}$$

$$= \frac{1}{2}\left(\sqrt{1 + \frac{4\eta}{\nu}} - 1\right) \tag{49}$$

$$= C_{\max}. \tag{50}$$

Using $C < C_{\max}$ and (50) ensures the existence of $\rho_0 \in \left(0, 2\left(\eta/\nu\right)^2\right)$ such that $w_{\rho_0,2} > C$. Moreover, the second degree coefficient of P_{ρ_0} is strictly positive, so that P_{ρ_0} is strictly negative on $(w_{\rho_0,1}, w_{\rho_0,2}) \supset [0, C]$ which completes the proof.

□

4 Asymptotical analysis of SABRINA

4.1 Stochastic majoration of $(F(\boldsymbol{x}_k))_{k\in\mathbb{N}}$

Proposition 1 *Under Assumptions 1-4, the following majoration holds almost surely:*

$$(\forall k \in \mathbb{N}) \quad \mathbb{E}\left[F(\mathbf{x}_{k+1})|\mathcal{F}_k\right] \leq F(\mathbf{x}_k) + \gamma_k \|\nabla F(\mathbf{x}_k)\|^2 P_{\gamma_k}(C), \tag{51}$$

where P_{γ_k} is the polynomial quantity defined in Lemma 3.

Proof Let $k \in \mathbb{N}$. We start by using the majoration property (2)-(3) of $h(., \mathbf{x}_k)$ on F at $\mathbf{x}_{k+1}$

$$F(\mathbf{x}_{k+1}) \leq F(\mathbf{x}_k) + \nabla F(\mathbf{x}_k)^\top (\mathbf{x}_{k+1} - \mathbf{x}_k) + \frac{1}{2}\|\boldsymbol{x}_{k+1} - \mathbf{x}_k\|^2_{\mathbf{A}_k}, \tag{52}$$

$$\leq F(\mathbf{x}_k) + \nabla F(\mathbf{x}_k)^\top (\mathbf{x}_{k+1} - \mathbf{x}_k) + \frac{\nu}{2}\|\mathbf{x}_{k+1} - \mathbf{x}_k\|^2 \quad a.s. \tag{53}$$

where (53) is a direct consequence of Assumption 2.

Using scheme (12) and the definition (8), inequality (53) can be written:

$$F(\mathbf{x}_{k+1}) \leq F(\mathbf{x}_k) - \gamma_k \nabla F(\mathbf{x}_k)^\top \mathbf{B}_k \mathbf{g}_k + \frac{\nu}{2}\|\mathbf{x}_{k+1} - \mathbf{x}_k\|^2, \tag{54}$$

$$= F(\mathbf{x}_k) - \gamma_k\, \mathbf{g}_k^\top \mathbf{B}_k \mathbf{g}_k + \gamma_k\, \mathbf{e}_k^\top \mathbf{B}_k \mathbf{g}_k + \frac{\nu}{2}\|\mathbf{x}_{k+1} - \mathbf{x}_k\|^2, \tag{55}$$

$$= F(\mathbf{x}_k) - \gamma_k\, \mathbf{g}_k^\top \mathbf{B}_k \mathbf{g}_k + \gamma_k\, \mathbf{e}_k^\top \mathbf{B}_k \nabla F(\mathbf{x}_k) + \gamma_k\, \mathbf{e}_k^\top \mathbf{B}_k \mathbf{e}_k + \frac{\nu}{2}\|\mathbf{x}_{k+1} - \mathbf{x}_k\|^2 \quad a.s. \tag{56}$$

On the one hand, Assumption 3(ii) guarantees that $\mathbf{g}_k \in \mathrm{Ker}(\mathbf{D}_k^\top)^\perp$ almost surely. Hence, the left inequality (17c) of Lemma 1 yields

$$\mathbf{g}_k^\top \mathbf{B}_k \mathbf{g}_k \geq \frac{1}{\nu}\|\mathbf{g}_k\|^2 \quad a.s. \tag{57}$$

On the other hand, the use of Cauchy-Schwarz inequality and relation (17b) from Lemma 1 gives

$$\mathbf{e}_k^\top \mathbf{B}_k \nabla F(\mathbf{x}_k) \leq \frac{1}{\eta} \|\nabla F(\mathbf{x}_k)\| \, \|\mathbf{e}_k\| \quad a.s. \tag{58}$$

Moreover, (17b) also leads to:

$$\mathbf{e}_k^\top \mathbf{B}_k \mathbf{e}_k \leq \frac{1}{\eta} \|\mathbf{e}_k\|^2 \quad a.s. \tag{59}$$

And, again as a consequence of (17b),

$$\|\mathbf{x}_{k+1} - \mathbf{x}_k\|^2 = \gamma_k^2 \|\mathbf{B}_k \mathbf{g}_k\|^2 \tag{60}$$

$$\leq \frac{\gamma_k^2}{\eta^2} \|\mathbf{g}_k\|^2 \quad a.s. \tag{61}$$

Plugging (57)-(61) into (56) leads to:

$$F(\mathbf{x}_{k+1}) \leq F(\mathbf{x}_k) - \frac{\gamma_k}{\nu} \|\mathbf{g}_k\|^2 + \frac{\gamma_k}{\eta} \|\nabla F(\mathbf{x}_k)\| \, \|\mathbf{e}_k\| + \frac{\gamma_k}{\eta} \|\mathbf{e}_k\|^2 + \frac{\nu \gamma_k^2}{2\eta^2} \|\mathbf{g}_k\|^2 \quad a.s. \tag{62}$$

Thanks to Lemma 2, we can take the conditional expectation in (62) and use the fact that it is a linear operator. Moreover, accounting for $\mathcal{F}_k$-measurability of $F(\mathbf{x}_k)$ and $\nabla F(\mathbf{x}_k)$, we obtain

$$\begin{aligned} \mathbb{E}\left[F(\mathbf{x}_{k+1})|\mathcal{F}_k\right] \leq F(\mathbf{x}_k) \; &- \frac{\gamma_k}{\nu} \, \mathbb{E}\left[\|\mathbf{g}_k\|^2 |\mathcal{F}_k\right] + \frac{\gamma_k}{\eta} \, \|\nabla F(\mathbf{x}_k)\| \, \mathbb{E}\left[\|\mathbf{e}_k\| \, |\mathcal{F}_k\right] \\ &+ \frac{\gamma_k}{\eta} \, \mathbb{E}\left[\|\mathbf{e}_k\|^2 |\mathcal{F}_k\right] + \frac{\nu \gamma_k^2}{2\eta^2} \, \mathbb{E}\left[\|\mathbf{g}_k\|^2 |\mathcal{F}_k\right] \quad a.s. \end{aligned} \tag{63}$$

The end of the proof aims at finding an upper bound of the last four terms in (63), depending only on $\nabla F(\mathbf{x}_k)$.

First, Definition (8) and the parallelogram identity give

$$\mathbb{E}\left[\|\mathbf{g}_k\|^2 |\mathcal{F}_k\right] = \|\nabla F(\mathbf{x}_k)\|^2 + 2 \, \mathbb{E}\left[\nabla F(\mathbf{x}_k)^\top \mathbf{e}_k |\mathcal{F}_k\right] + \mathbb{E}\left[\|\mathbf{e}_k\|^2 |\mathcal{F}_k\right] \quad a.s. \tag{64}$$

Since $\nabla F(\boldsymbol{x}_k)$ is $\mathcal{F}_k$-measurable, and using Assumption 4(i), we have

$$\mathbb{E}\left[\nabla F(\mathbf{x}_k)^\top \mathbf{e}_k |\mathcal{F}_k\right] = \nabla F(\mathbf{x}_k)^\top \mathbb{E}\left[\mathbf{e}_k |\mathcal{F}_k\right] \tag{65}$$

$$= 0 \quad a.s, \tag{66}$$

which leads to the conditional equality

$$\mathbb{E}\left[\|\mathbf{g}_k\|^2 |\mathcal{F}_k\right] = \|\nabla F(\mathbf{x}_k)\|^2 + \mathbb{E}\left[\|\mathbf{e}_k\|^2 |\mathcal{F}_k\right] \quad a.s. \tag{67}$$

Using Assumption 4(ii) we then deduce the following bounds for $\|\mathbf{g}_k\|^2$

$$\|\nabla F(\mathbf{x}_k)\|^2 \leq \; \mathbb{E}\left[\|\mathbf{g}_k\|^2 |\mathcal{F}_k\right] \leq \left(1 + C^2\right) \|\nabla F(\mathbf{x}_k)\|^2 \quad a.s. \tag{68}$$

Second, the following stochastic majoration of $\mathbb{E}\left[\|\boldsymbol{\epsilon}_k\| \ |\mathcal{F}_k\right]$ is obtained by Jensen's inequality and Equation (16)

$$\mathbb{E}\left[\|\mathbf{e}_k\| \ |\mathcal{F}_k\right] \leq \sqrt{\mathbb{E}\left[\|\mathbf{e}_k\|^2|\mathcal{F}_k\right]} \tag{69}$$
$$\leq C\|\nabla F(\mathbf{x}_k)\| \quad a.s. \tag{70}$$

where (70) arises from Assumption 4(ii).

Finally, Inequalities (16), (68), (70) combined with (63) give the desired result

$$\mathbb{E}\left[F(\mathbf{x}_{k+1})|\mathcal{F}_k\right] \leq F(\mathbf{x}_k)+\gamma_k\|\nabla F(\mathbf{x}_k)\|^2\left[\left(1+\frac{\nu\gamma_k}{2\eta}\right)\frac{C^2}{\eta}+\frac{C}{\eta}+\left(\frac{\nu\gamma_k}{2\eta^2}-\frac{1}{\nu}\right)\right] \quad a.s. \tag{71}$$

□

Proposition 2 *Under Assumptions 1-5, for every $\rho > 0$, there exists k_ρ such that*

$$(\forall k \geq k_\rho) \quad \mathbb{E}\left[F(\mathbf{x}_{k+1})|\mathcal{F}_k\right] \leq F(\mathbf{x}_k) + \gamma_k\|\nabla F(\mathbf{x}_k)\|^2 P_\rho(C) \quad a.s. \tag{72}$$

Proof By Assumption 5, $\gamma_k \underset{k\to+\infty}{\longrightarrow} 0$, which ensures the existence of k_ρ such that $\gamma_k \leq \rho$ for all $k \geq k_\rho$. Thus,

$$P_{\gamma_k}(C) = \left[\left(1+\frac{\nu\gamma_k}{2\eta}\right)\frac{C^2}{\eta}+\frac{C}{\eta}+\left(\frac{\nu\gamma_k}{2\eta^2}-\frac{1}{\nu}\right)\right],$$
$$\leq \left[\left(1+\frac{\nu\rho}{2\eta}\right)\frac{C^2}{\eta}+\frac{C}{\eta}+\left(\frac{\nu\rho}{2\eta^2}-\frac{1}{\nu}\right)\right] = P_\rho(C). \tag{73}$$

Inequality (72) directly follows from (51) of Proposition 1.

□

4.2 General convergence theorem

We start with the following theorem which gives a general result for SABRINA without any convexity hypothesis:

Theorem 1 *Under Assumptions 1-5, sequence $(F(\mathbf{x}_k))_{k\in\mathbb{N}}$ converges a.s. to an almost surely finite random variable. Moreover, $(\mathbf{x}_k)_{k\in\mathbb{N}}$ is such that*

$$\sum_{k=0}^{+\infty}\|\mathbf{x}_{k+1}-\mathbf{x}_k\|^2 < +\infty \quad a.s., \tag{74a}$$
$$\liminf_{k\to+\infty}\|\nabla F(\mathbf{x}_k)\| = 0 \quad a.s. \tag{74b}$$

Proof From Lemma 3, there exists $\rho_0 > 0$ for which P_{ρ_0} is strictly negative on $[0, C]$. Applying Proposition 2 with $\rho = \rho_0$, yields the existence of k_{ρ_0} such that

$$(\forall k \geq k_{\rho_0}) \quad \mathbb{E}\left[F(\mathbf{x}_{k+1})|\mathcal{F}_k\right] \leq F(\boldsymbol{x}_k) + \gamma_k\|\nabla F(\mathbf{x}_k)\|^2 P_{\rho_0}(C) \quad a.s. \tag{75}$$

Subtracting F^*, the minimal value of F on each side of (75) yields

$$(\forall k \geq k_{\rho_0}) \quad \mathbb{E}\left[F(\mathbf{x}_{k+1})-F^*|\mathcal{F}_k\right] \leq [F(\mathbf{x}_k)-F^*] + \gamma_k\|\nabla F(\mathbf{x}_k)\|^2 P_{\rho_0}(C) \quad a.s. \tag{76}$$

All random variables involved in (76) are positive and integrable. Moreover, we have $P_{\rho_0}(C) < 0$ (since P_{ρ_0} is strictly negative on $[0, C]$). Thus, we can invoke Robbins-Siegmund's lemma [62]. The *a.s.* convergence of $(F(\mathbf{x}_k) - F^*)_{k\in\mathbb{N}}$ to an *a.s.* finite random variable is guaranteed, and so it is for $(F(\mathbf{x}_k))_{k\in\mathbb{N}}$. Moreover again from Robbins-Siegmund's lemma, we have the following property

$$\sum_{k=0}^{+\infty} \gamma_k \|\nabla F(\mathbf{x}_k)\|^2 < +\infty \quad a.s. \tag{77}$$

First, using (12), (17b) and then (68), yields

$$\sum_{k=0}^{+\infty} \mathbb{E}\left[\|\mathbf{x}_{k+1} - \mathbf{x}_k\|^2 | \mathcal{F}_k\right] \leq \sum_{k=0}^{+\infty} \frac{\gamma_k^2}{\eta^2}\ E[\|\mathbf{g}_k\|^2 | \mathcal{F}_k] \tag{78}$$

$$\leq \frac{1 + C^2}{\eta^2} \sum_{k=0}^{+\infty} \gamma_k^2 \|\nabla F(\mathbf{x}_k)\|^2 \quad a.s. \tag{79}$$

By Assumption 5, $(\gamma_k)_{k\in\mathbb{N}}$ is positive and converges to 0. Thus, $\gamma_k^2 \|\nabla F(\mathbf{x}_k)\|^2 \leq \gamma_k \|\nabla F(\mathbf{x}_k)\|^2$ from a certain range k. It follows that the right term in (79) is a finite random variable and, as a consequence, $\sum_{k=0}^{+\infty} \mathbb{E}\left[\|\mathbf{x}_{k+1} - \mathbf{x}_k\|^2 | \mathcal{F}_k\right] < +\infty$. Positivity of sequence $\left(\|\mathbf{x}_{k+1} - \mathbf{x}_k\|^2\right)_{k\in\mathbb{N}}$ finally allows us to apply [52, Ch.1, Th. 21] which gives (74a).

Our proof of (74b) is similar to the one of Zoutendijk condition for gradient-based optimization methods [10], adapted to a stochastic framework. To do so, we stand on complementary set $\left\{\omega \in \Omega \,|\, \liminf_{k\to+\infty} \|\nabla F(\mathbf{x}_k(\omega))\| > 0\right\}$ and prove that it is of zero probability.

For all $\omega \in \Omega$ such that $\liminf_{k\to+\infty} \|\nabla F(\mathbf{x}_k(\omega))\| > 0$, following the definition of lim inf, there exists $\varepsilon(\omega) > 0$ and a range $\mathbf{k}_0(\omega) \in \mathbb{N}$ for which for all $k \geq \mathbf{k}_0(\omega)$, $\|\nabla F(\mathbf{x}_k(\omega))\| \geq \varepsilon(\omega)$. Thus

$$(\forall k \geq \mathbf{k}_0(\omega)) \quad \gamma_k \|\nabla F(\mathbf{x}_k(\omega))\|^2 \geq \gamma_k \varepsilon(\omega)^2. \tag{80}$$

Summing (80) from $\mathbf{k}_0(\omega)$ to $+\infty$, and using Assumption 5, we deduce

$$\sum_{k=\mathbf{k}_0(\omega)}^{+\infty} \gamma_k \|\nabla F(\mathbf{x}_k(\omega))\|^2 \geq \varepsilon(\omega)^2 \sum_{k=\mathbf{k}_0(\omega)}^{+\infty} \gamma_k \tag{81}$$

$$= +\infty. \tag{82}$$

This leads to inclusion

$$\left\{\omega \in \Omega \,|\, \liminf_{k\to+\infty} \|\nabla F(\mathbf{x}_k(\omega))\| > 0\right\} \subset \left\{\omega \in \Omega \,|\, \sum_{k=0}^{+\infty} \gamma_k \|\nabla F(\mathbf{x}_k(\omega))\|^2 = +\infty\right\}. \tag{83}$$

The term in the right side of (83) is a negligible set according to (77). As a consequence, the left side of (83) is also a negligible set and (74b) holds by taking the complement.

□

Although some recent works consider (74b) as a sufficient convergence criterion [42], its scope remains limited since it only holds for a given subsequence of $(\mathbf{x}_k)_{k\in\mathbb{N}}$. In the following, we make use of topological arguments to derive useful corollaries of Theorem 1.

Corollary 1 *Under Assumptions 1-5, there exists a full measure subset Λ such that, for every $\omega \in \Lambda$, the following statements hold:*

(i) $(F(\mathbf{x}_k(\omega))_{k\in\mathbb{N}}$ *converges to a finite limit ;*
(ii) $(\mathbf{x}_k(\omega))_{k\in\mathbb{N}}$ *is bounded ;*
(iii) $\chi^\infty(\omega)$*, the set of accumulation points of* $(\mathbf{x}_k(\omega))_{k\in\mathbb{N}}$*, is non empty, compact, connex and contains at least one element of* $\operatorname{zer}\nabla F$*.*

Proof Since Theorem 1 holds *a.s.*, there exists a set $\Lambda \subset \Omega$ of probability one where, for all $\omega \in \Lambda$,

$$\lim_{k\to+\infty} F(\mathbf{x}_k(\omega)) < +\infty, \tag{84a}$$

$$\sum_{k=0}^{+\infty} \|\mathbf{x}_{k+1}(\omega) - \mathbf{x}_k(\omega)\|^2 < +\infty \tag{84b}$$

$$\liminf_{k\to+\infty} \|\nabla F(\mathbf{x}_k(\omega))\| = 0. \tag{84c}$$

Inequality (84a) implies that $(F(\mathbf{x}_k(\omega)))_{k\in\mathbb{N}}$ is a bounded sequence. The coercivity of F, in Assumption 1, ensures this same property for $(\mathbf{x}_k(\omega))_{k\in\mathbb{N}}$. It follows that the set of cluster points $\chi^\infty(\omega)$ is non empty and bounded. Moreover, it is compact due to its closure (in finite dimension).

Moreover, (84b) leads to:

$$\mathbf{x}_{k+1}(\omega) - \mathbf{x}_k(\omega) \underset{k\to+\infty}{\longrightarrow} \mathbf{0}_N. \tag{85}$$

Equation (85), and the boundedness of $(\mathbf{x}_k(\omega))_{k\in\mathbb{N}}$ enables the use of Ostrowski's theorem [3] [57, Th. 26.1]) which directly gives the connexity of $\chi^\infty(w)$.

From the boundedness of $(\mathbf{x}_k(\omega))_{k\in\mathbb{N}}$, and (84c), we deduce that there exists a convergent sub-sequence $\left(\mathbf{x}_{\varphi(k)}(\omega)\right)_{k\in\mathbb{N}}$ such that

$$\nabla F(\mathbf{x}_{\varphi(k)}(\omega)) \underset{k\to+\infty}{\longrightarrow} \mathbf{0}_N. \tag{86}$$

Let us denote by $\mathbf{x}^\infty(w)$ the limit point of $\left(\mathbf{x}_{\varphi(k)}(\omega)\right)_k$. By construction, $\mathbf{x}^\infty(\omega)$ belongs to $\chi^\infty(\omega)$. Since F is gradient-Lipschitz, by Assumption 1, its gradient is continuous and we finally obtain:

$$\nabla F(\mathbf{x}^\infty(\omega)) = \mathbf{0}_N. \tag{87}$$

□

[3] Let $(\boldsymbol{z}_k)_{k\in\mathbb{N}}$ a bounded sequence of $\mathbb{R}^N$, verifying $\boldsymbol{z}_{k+1} - \boldsymbol{z}_k \underset{k\to+\infty}{\longrightarrow} 0$. Then the set of cluster points of $(\boldsymbol{z}_k)_{k\in\mathbb{N}}$ is connex.

Corollary 1 provides us an overview of the distribution formed by the accumulation points of sequence $(\mathbf{x}_k)_{k\in\mathbb{N}}$. In order to refine the convergence theorem, we must introduce extra assumptions on the level sets of function F (see 2.1 for useful notations). From this perspective, we propose a result, when F is convex with isolated stationary points.

Corollary 2 *Under Assumptions 1-5, if F is convex with isolated stationary points. Then $(\mathbf{x}_k)_{k\in\mathbb{N}}$ converges almost surely to the (global) minimizer of F.*

Proof Convexity of F ensures that $\operatorname{zer}\nabla F$ matches with the set of minimizers of F, and that this set is convex. Thus, if the stationary points of F are isolated, F admits a unique minimizer $\boldsymbol{x}^*$ and then $\operatorname{zer}\nabla F = \{\boldsymbol{x}^*\}$.

Let us now consider the probability set Λ from Corollary 1 and some $\omega \in \Lambda$. Sequence $(\mathbf{x}_k(\omega))_{k\in\mathbb{N}}$ possesses a cluster point which lies in $\operatorname{zer}\nabla F$. Since $\operatorname{zer}\nabla F = \{\boldsymbol{x}^*\}$, it follows that $\boldsymbol{x}^* \in \chi^\infty(\omega)$. Moreover, F is continuous and $F(\mathbf{x}_k(\omega))_{k\in\mathbb{N}}$ converges to a finite limit. Thus, $F(\mathbf{x}_k(\omega)) \underset{k\to+\infty}{\longrightarrow} F(\boldsymbol{x}^*)$. It follows that every $\tilde{\boldsymbol{x}} \in \chi^\infty(\omega)$ also verifies $F(\tilde{\boldsymbol{x}}) = F(\boldsymbol{x}^*)$ and thus is a minimizer of F. The uniqueness of the latter gives $\chi^\infty(\omega) = \{\boldsymbol{x}^*\}$. Finally, the boundedness of the sequence $(\mathbf{x}_k(\omega))_{k\in\mathbb{N}}$ ensures its convergence to $\boldsymbol{x}^*$.

The fact that Λ is a set of probability one concludes the proof.

4.3 Convergence rate analysis

We provide here our second main theoretical result, regarding the convergence rate of SABRINA, in the case when F satisfies a strong convexity property.

Theorem 2 *If F is α-strongly convex (i.e., $F - \frac{\alpha}{2}\|\cdot\|^2$ is convex) and Lipschitz differentiable function on $\mathbb{R}^N$ then, under Assumptions 2-5, there exists a sequence $(r_k)_{k\in\mathbb{N}}$ such that, for k sufficiently large,*

$$\mathbb{E}\left[F(\mathbf{x}_{k+1}) - F^*\right] \leq e^{r_k}, \tag{88}$$

where

$$r_k \underset{k\to+\infty}{\sim} 2\alpha\,|P_{\rho_0}(C)| \times \left(-\sum_{i=0}^{k}\gamma_i\right), \tag{89}$$

with P_{ρ_0} the polynomial defined in Lemma 3.

Proof First, Theorem 2 assumes that F is supposed to be α-convex and Lipschitz differentiable. Thus, Assumption 1 holds. Since Assumptions 2-5 also hold, we can thus come back to (76) (from the proof of Theorem 1, and using the same notations), and take the expectation to obtain:

$$(\forall k \geq k_{\rho_0}) \quad \mathbb{E}\left[F(\mathbf{x}_{k+1}) - F^*\right] \leq \mathbb{E}\left[F(\mathbf{x}_k) - F^*\right] + \gamma_k P_{\rho_0}(C)\,\mathbb{E}\left[\|\nabla F(\mathbf{x}_k)\|^2\right] \quad a.s. \tag{90}$$

Let us make use of [13, Eq. (4.12)] related to strongly convex functions, which reads:

$$(\forall k \in \mathbb{N}) \quad \|\nabla F(\mathbf{x}_k)\|^2 \geq 2\alpha(F(\mathbf{x}_k) - F^*). \tag{91}$$

Substituting (91) in (90) then leads to:

$$(\forall k \geq k_{\rho_0}) \quad \mathbb{E}\left[F(\mathbf{x}_{k+1}) - F^*\right] \leq (1 + \widehat{\gamma}_k)\, \mathbb{E}\left[F(\mathbf{x}_k) - F^*\right], \tag{92}$$

with

$$\widehat{\gamma}_k = 2\alpha P_{\rho_0}(C)\gamma_k \tag{93}$$

$$< 0 \quad (\text{since} \quad P_{\rho_0}(C) < 0). \tag{94}$$

Moreover, by Assumption 5, $(\gamma_k)_{k\in\mathbb{N}}$ converges to 0 so that there exists $k_1 > k_{\rho_0}$ such that:

$$(\forall k \geq k_1) \quad 1 + \widehat{\gamma}_k \in (0, 1). \tag{95}$$

Then, by induction, it follows that for all $k \geq k_1 + 1$,

$$\mathbb{E}\left[F(\mathbf{x}_k) - F^*\right] \leq \mathbb{E}\left[F\left(\mathbf{x}_{k_1}\right) - F^*\right] \prod_{i=k_1}^{k-1} (1 + \widehat{\gamma}_i)\,. \tag{96}$$

Taking the logarithm in (96), by virtue of Condition (95), then yields:

$$\ln\left(\mathbb{E}\left[F(\mathbf{x}_k) - F^*\right]\right) \leq \sum_{i=k_1}^{k-1} \ln\left(1 + \widehat{\gamma}_i\right) + \ln\left(\mathbb{E}\left[F\left(\mathbf{x}_{k_1}\right) - F^*\right]\right). \tag{97}$$

The end of the proof consists in searching for an asymptotic equivalent of the right member of (97). Convergence of $(\gamma_k)_{k\in\mathbb{N}}$ to 0 (by Assumption 5) ensures:

$$\ln(1 + \widehat{\gamma}_k) \underset{k\to+\infty}{\sim} \widehat{\gamma}_k. \tag{98}$$

Sequences $\left(\ln(1 + \widehat{\gamma}_k)\right)_{k\geq k_1}, \left(\widehat{\gamma}_k\right)_{k\geq k_1}$ are both negative. Moreover, Assumption 5 yields:

$$\sum_{i=k_1}^{+\infty} \widehat{\gamma}_i = -\infty. \tag{99}$$

We can thus deduce:

$$\sum_{i=k_1}^{k-1} \ln(1 + \widehat{\gamma}_i) \underset{k\to+\infty}{\sim} \sum_{i=k_1}^{k-1} \widehat{\gamma}_i, \tag{100}$$

$$= 2\alpha P_{\rho_0}(C) \sum_{i=k_1}^{k-1} \gamma_i. \tag{101}$$

Since the series $\sum_{i=k_1}^{k-1} \ln(1 + \widehat{\gamma}_i)$ diverges to $-\infty$, it follows that

$$\sum_{i=k_1}^{k-1} \ln\left(1 + \widehat{\gamma}_i\right) + \ln\left(\mathbb{E}\left[F\left(\boldsymbol{x}_{k_1}\right) - F^*\right]\right) \underset{k\to+\infty}{\sim} 2\alpha P_{\rho_0}(C) \sum_{i=k_1}^{k-1} \gamma_i \tag{102}$$

$$\underset{k\to+\infty}{\sim} 2\alpha P_{\rho_0}(C) \sum_{i=0}^{k} \gamma_i. \tag{103}$$

Going from (102) to (103) arises from Assumption 5. The desired conclusion is reached passing to the exponential.

□

Since $\sum_{i=0}^{k} \gamma_k = +\infty$ (by Assumption 5), Theorem 2 guarantees the L^1 convergence to F^* for sequence $F(\mathbf{x}_k)_{k\in\mathbb{N}}$ generated by SABRINA. It should be emphasized that Assumption 5 is rather mild. One interesting practical choice consists in setting $(\gamma_k)_{k\in\mathbb{N}}$ as a sequence converging to zero as slow as allowed by Assumption 5. As a result, relation (89) ensures that the logarithmic expectation in (88) converges fast to minus infinity.

4.4 Link to existing works

Our "liminf" convergence criterion (74b) is probably the most encountered one in optimization [10, Ch. 1.4] among those introduced in Theorem 1. Similar result is also obtained in [39, 42] considering a stochastic context. The aforementioned works focused on an method close to ADAM [45], that has been quite notorious in the field of deep learning this last decade. To a certain extent, the scheme in [39, 42] can be interpreted as a specific case of ours without using MM metric (i.e., $\mathbf{A}_k \equiv \boldsymbol{I}_N$) and where subspace acceleration is replaced by momentum weights combining with a manually tuned stepsize. By including an MM approach in a non-convex situation, the MISO algorithm from [50] shares common features with the one we develop here. The asymptotical result from [50, Prop. 3.3] is also expressed as a "liminf" condition but, up to our knowledge, might appear harder to interpret than (74b). Our result (74a) is not as common in the litterature of stochastic optimization, as its counterpart (74b), probably since it is slightly less tractable. It shares structural similarities with the finite length condition stated in [23, Th. 3] studying the MM subspace algorithm without noisy gradient. When considering noisy gradient, we manage here to show (74a), which is weaker in the sense that the square norm summation does not ensure necessarily that $(\mathbf{x}_k)_{k\in\mathbb{N}}$ is a Cauchy sequence, and thus does not allow to easily conclude on its almost sure convergence.

More generally, Robbins-Siegmund's lemma [62] is a widely used tool to deduce asymptotical properties of stochastic approximation schemes [31, 38]. Our use of Ostrowski's theorem [57, Th. 26.1] and the connexity argument to obtain Corollary 1 is reminiscent from [25]. However, in contrast with the aforementioned work, we use the convexity hypothesis only at the very end in Corollary 2. The idea of using level-set as an alternative arises from [38].

The supervised learning context has highly promoted studies relative to speed estimation of stochastic algorithms and especially for gradient methods both in a convex [13, 38, 55] and more recently in a non-convex setting [34]. [11, 16] also focus on quasi-Newton approximation approaches (and obtain an L^1 convergence result). These methods are actually similar to SABRINA, for a particular subspace choice. However, in contrast with our approach, no MM metric/stepsize is used in [11, 16].

5 Application to binary classification

As a first illustrative example, we focus on a supervised binary classification problem. We consider M feature vectors $(\boldsymbol{v}_m)_{1\leq m\leq M} \in \mathbb{R}^N$, with their associated labels $(y_m)_{1\leq m\leq M} \in \{-1,1\}$ as a training dataset. In a linear classification context, one possibility to estimate the parameter model $\boldsymbol{x}^* \in \mathbb{R}^N$ consists in searching the best linear classifier through the minimization of the log-loss penalized empirical risk [15]:

$$(\forall \boldsymbol{x} \in \mathbb{R}^N) \quad F(\boldsymbol{x}) = \frac{1}{M}\sum_{m=1}^{M} \log\left(1+\exp\left(-[\boldsymbol{H}\boldsymbol{x}]_m\right)\right) + \mu \sum_{n=1}^{N} \log\left(1+\frac{x_n^2}{\delta^2}\right). \tag{104}$$

Matrix $\boldsymbol{H} = \mathrm{Diag}\{(y_m)_{1\leq i\leq M}\}[\boldsymbol{v}_1,...,\boldsymbol{v}_M]^\top \in \mathbb{R}^{M\times N}$ involved in the so-called data-fidelity term, gathers the information brought by the training dataset. The second term in (104) is a regularization term weighted by $\mu > 0$, which aims at promoting the sparsity of the estimated model so as to limit overfitting issues. The retained regularization is a coercive, continuous but non-convex approximation of the ℓ_0 norm, which is at the core of re-weighted ℓ_1 schemes [18,59]. Function (104) is Lipschitz differentiable on $\mathbb{R}^N$ and coercive. However, it is non convex due to the regularization term.

5.1 Majorant mapping and convergence guarantees

Let us to build a majorant mapping for the objective function (104). Thanks to the additivity of the majoration property (see, for eg, [69]), we can majorize each term separately. A majorant mapping for the log-loss can be deduced from [14, Eq.5], while a majorant mapping for the non-convex penalty was provided in [25, Tab.I]. This yields the following majorant mapping $\boldsymbol{A}(\cdot)$ for the objective function (104):

$$\begin{aligned}(\forall \boldsymbol{x} \in \mathbb{R}^N) \quad & \boldsymbol{A}(\boldsymbol{x}) \\ & = \boldsymbol{H}^\top \mathrm{Diag}\left\{(\vartheta([\boldsymbol{L}\boldsymbol{x}]_m)_{1\leq m\leq M}\right\}\boldsymbol{H} + \mu \mathrm{Diag}\left\{\left(\frac{2}{x_n^2+\delta^2}\right)_{1\leq n\leq N}\right\} + \tau \boldsymbol{I}_N,\end{aligned} \tag{105}$$

with $\vartheta : u \mapsto \frac{1}{u}\left(\frac{1}{1+\exp(-u)} - \frac{1}{2}\right)$ extended by continuity in 0. Moreover, τ is a strictly positive constant ensuring the fulfilment of Assumption 2. For such choice of mapping, Assumption 2 holds with:

$$\eta = \tau, \quad \nu = \tau + \frac{1}{4M}|||\boldsymbol{H}|||^2 + 2\frac{\mu}{\delta^2}. \tag{106}$$

We propose to implement SABRINA by considering two choices for the subspace, namely $\mathbf{D}_k = \boldsymbol{I}_N$, and $\mathbf{D}_k = [-\mathbf{g}_k \mid \mathbf{x}_k - \mathbf{x}_{k-1}]$. For the latter, the second column of $\mathbf{D}_k$ is removed if the rank of $\mathbf{D}_k$ gets lower than $M_k = 2$. Note that this situation never happened in our practical experiments. Both subspaces thus satisfy Assumption 3 and respectively yield the so-called SABRINA-I and SABRINA-MG algorithms. If Assumptions 4 and 5 hold, sequences generated by these two

algorithms verify Theorem 1 and Corollary 1. Otherwise stated, for suitable step-size and noise perturbation settings, our theoretical analysis ensures an almost sure convergence to a stationary point of F for a subsequence of $(\mathbf{x}_k)_{k\in\mathbb{N}}$. Function F is non convex and does not have finite level sets, so that the stronger convergence results established in our study cannot be applied.

5.2 Numerical settings

When using the SABRINA-I scheme, the majorant function minimization requires to invert an $N \times N$ system, which is performed using the linear solver from [68]. The gradient perturbation is simulated by applying a multiplicative noise following a uniform law centered in 0 on every component of the gradient at each iteration that is, for every $k \in \mathbb{N}$,

$$\mathbf{e}_k = C \times \mathrm{Diag}\{(\mathrm{u}_{n,k})_{1\leq n\leq N}\}\nabla F(\mathbf{x}_k), \tag{107}$$

where each entry of $\mathbf{u}_k = (\mathrm{u}_{n,k})_{1\leq n\leq N} \in \mathbb{R}^N$ is an independant realization of a uniform law between $[-1, 1]$. By construction, Condition (16) holds since, for every $k \in \mathbb{N}$:

$$\mathbb{E}\left[\|\mathbf{e}_k\|^2|\mathcal{F}_k\right] \leq C^2\mathbb{E}\left[\|\mathrm{Diag}\{(\mathrm{u}_{n,k})_{1\leq n\leq N}\}\mathbf{u}_k\|_\infty|\mathcal{F}_k\right]\|\nabla F(\mathbf{x}_k)\|^2 \tag{108}$$

$$= C^2\|\nabla F(\mathbf{x}_k)\|^2. \tag{109}$$

Equation (108) also guarantees the integrability of $\mathbf{e}_k$. Moreover, $\mathbf{u}_k$ is zero-mean so that Assumption 4(i) also holds. We set the decreasing step-size $\gamma_k = 1/(k+1)^{0.01}$, for $k \in \mathbb{N}$, thus satisfying Assumption 5. Performance of SABRINA are evaluated against those of state-of-the-art stochastic gradient-based schemes from the machine learning field, namely SGD [7], ADAGRAD [29] and RMSprop [71]. The parameter tuning for these methods (e.g., learning rate, momentum weight) was made empirically, following recommendations from [65], to obtain best possible practical convergence behaviours.

Datasets `rcv1` and `a8a` are extracted from LIBSVM library [21]. Table 1 lists properties of these datasets and the retained hyperparameters μ, δ and τ. The latter has been manually chosen to ensure a satisfying compromise between a good conditioning of the majorant mapping (and then, a wide range of values for C, see Sec. 3.3) and a fast convergence rate.

Dataset	Train M	Test	Features N	$\|\|\|\boldsymbol{H}\|\|\|^2/(4M)$	μ	δ	τ	$C_{\max}$
`rcv1`	20242	677399	47236	5.5×10^{-3}	10^{-1}	1	1	0.54
`a8a`	9865	22696	122	1.6	10^{-2}	1	0.5	0.2

Table 1: Dataset properties and hyperparameter settings

5.3 Experimental results

In Figs. 1 and 2, we illustrate the efficiency of every competitor through the evolution of the objective function, and of the gradient norm of their iterates along

time for a Matlab 2020a code ran on a desktop computer equipped with an Intel Core i7 3.2 GHz pro and 16 GB RAM. In Fig. 1, F^* (with a slight abuse of notation) denotes the function value computed numerically after running 3MG [23] method (i.e., SABRINA-MG without noise in the gradient), for a large number of iterations. For both figures, we set $C = 0.95 \times C_{\max}$, so as to meet the conditions imposed by Assumption 4(ii) and then convergence of SABRINA is ensured in the sense of Theorem 1 and Corollary 1. It is noticeable that both SABRINA variants reach the best performance when compared to their competitors. Moreover, for both datasets, the interest of subspace acceleration is visible, as SABRINA-MG reaches faster convergence than SABRINA-I. Finally, let us emphasize that SABRINA implementation does not impose any tedious manual learning/momentum rate tuning, as it was the case for the other methods. Table 2 lists classification scores obtain by SABRINA-MG at convergence for both datasets.

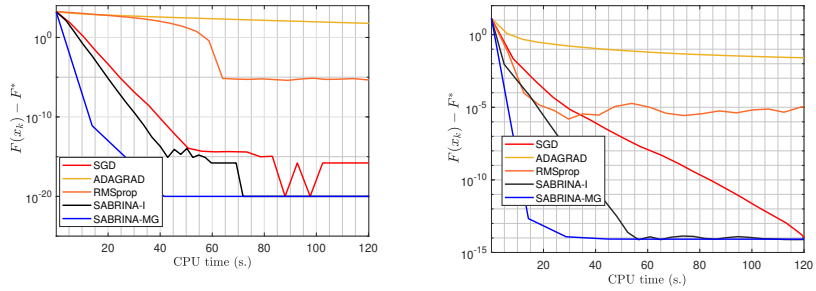

Fig. 1: Evolution of the objective function along time for various algorithms, on dataset `rcv1` (left) and `a8a` (right). Noise amplitude $C = 0.95 \times C_{\max}$.

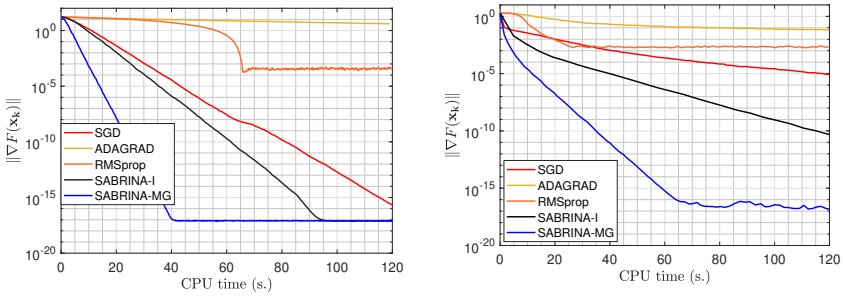

Fig. 2: Evolution of the gradient norm along time for various algorithms, on dataset `rcv1` (left) and `a8a` (right). Noise amplitude $C = 0.95 \times C_{\max}$.

Dataset	Accuracy	AUC	Precision	Recall
`rcv1`	$9,2 \times 10^{-1}$	$9,7 \times 10^{-1}$	$9,3 \times 10^{-1}$	$9,1 \times 10^{-1}$
`a8a`	$8,4 \times 10^{-1}$	$8,9 \times 10^{-1}$	$7,5 \times 10^{-1}$	$5,2 \times 10^{-1}$

Table 2: Classification scores after running SABRINA-MG for 60 s.

Fig. 3 illustrates the evolution of the gradient norm along SABRINA-MG iterations for various levels of noise on the gradient term, when considering the `rcv1` example. Increasing the noise level obviously slows down the convergence of the method. Moreover, one can see that SABRINA-MG starts showing some oscillating behaviour when $C \geq C_{\max}$. Considering an order of magnitude ten times higher than $C_{\max}$, one can observe a change of regime where the convergence of the algorithm seems compromised. Such phenomenons suggest that the bound $C_{\max}$ involved in Assumption 4 is consistent and not over pessimistic in this example for ensuring practical stability of the algorithm.

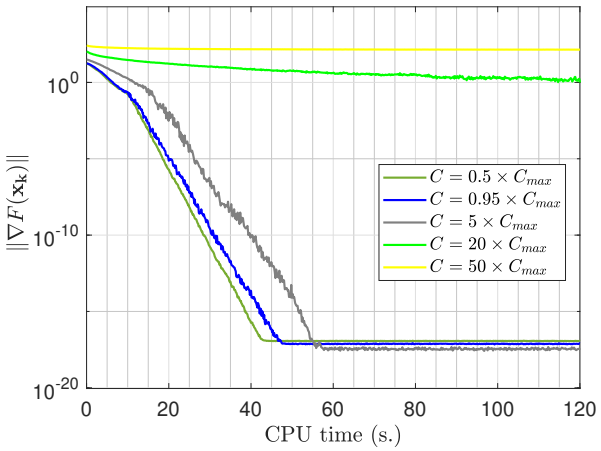

Fig. 3: `rcv1`: Evolution of the gradient norm along time for various noise amplitudes affecting the gradient term in SABRINA-MG.

6 Application to robust blur kernel identification

We now consider an inverse problem of robust blur kernel identification. The observation model is similar than in [25], namely

$$\boldsymbol{y} = B(\overline{\boldsymbol{x}})\boldsymbol{z} + \boldsymbol{n}, \tag{110}$$

where $\boldsymbol{z} \in \mathbb{R}^M$ and $\boldsymbol{y} \in \mathbb{R}^M$ are original and (blurry and noisy) degraded versions of a given image with M pixels, $\overline{\boldsymbol{x}} \in \mathbb{R}^N$ is an unknown blur kernel to be estimated, and $\boldsymbol{n} \in \mathbb{R}^M$ models additive noise. The blur operation corresponds to a 2D discrete convolution (with circulant-padding assumption) between $\boldsymbol{z}$ and $\overline{\boldsymbol{x}}$, represented by the linear operator $B : \mathbb{R}^N \to \mathbb{R}^M$. The goal is to retrieve an estimation of $\overline{\boldsymbol{x}}$ from the pair of images $(\boldsymbol{z}, \boldsymbol{y})$. This inverse problem typically arises

in the calibration of optical instruments [5, 43]. The observation model (110) can be expressed equivalently as

$$\boldsymbol{y} = \boldsymbol{H}\overline{\boldsymbol{x}} + \boldsymbol{n}, \tag{111}$$

where the blur operation is rewritten as the application of the linear Hankel-block operator $\boldsymbol{H} \in \mathbb{R}^{M\times N}$ (related to $\boldsymbol{z}$) on the kernel $\overline{\boldsymbol{x}}$. In contrast with [25], we consider the challenging noise scenario where outliers can arise in the observed data. Specifically, $\boldsymbol{n} \in \mathbb{R}^N$ is the realization of a Gaussian mixture noise with standard deviations $(\sigma_1, \sigma_2) > 0$ and mixing rate $\varrho \in]0, 1[$, where typically $\sigma_1 \ll \sigma_2$. An efficient strategy for solving (111) consists in minimizing a penalized criterion:

$$(\forall \boldsymbol{x} \in \mathbb{R}^N) \quad F(\boldsymbol{x}) = L(\boldsymbol{x}) + R(\boldsymbol{x}), \tag{112}$$

where L plays the role of the data fidelity term, accounting for the mixture noise model, and R is a regularization function promoting desirable prior assumption on the sought $\boldsymbol{x}$.

Due to the presence of outliers in the noise, we opt for the following Huber data fidelity term, well suited for robust inverse problem resolution,

$$(\forall \boldsymbol{x} \in \mathbb{R}^N) \quad L(\boldsymbol{x}) = \sum_{m=1}^{M} \ell_m([\boldsymbol{H}\boldsymbol{x}]_m]), \tag{113}$$

where

$$(\forall m = 1, \ldots, M)(\forall t \in \mathbb{R}) \quad \ell_m(t) = \begin{cases} \frac{1}{2}(t - y_m)^2 & \text{if } |t - y_m| \leq p \\ p|t - y_m| - \frac{1}{2}p^2 & \text{otherwise,} \end{cases} \tag{114}$$

with $p > 0$ some outlier threshold constant. Moreover, we choose to promote smoothness of the restored kernel, by setting:

$$(\forall \boldsymbol{x} \in \mathbb{R}^N) \quad R(\boldsymbol{x}) = \sum_{n=1}^{N} \psi(\|\boldsymbol{\Delta}_n \boldsymbol{x}\|). \tag{115}$$

Hereabove, for every $n \in \{1, \ldots, N\}$, $\boldsymbol{\Delta}_n \in \mathbb{R}^{2\times N}$ corresponds to the discrete vertical and horizontal gradient operators applied to the n-th pixel of the 2D reshaped kernel $\boldsymbol{x}$. Moreover, $\psi : u \mapsto \lambda\sqrt{1 + u^2/\kappa^2}$ is the hyperbolic penalty with smoothness parameter $\kappa > 0$. Function (115) can thus be viewed as a smoothed version of the classical total-variation norm widely used in image processing. Parameter $\lambda > 0$ is a regularization parameter.

The resulting function (112) is convex and Lipschitz differentiable on $\mathbb{R}^N$. Moreover, according to [63, Proposition 2.5], F is coercive if and only if

$$\mathrm{Ker}(\boldsymbol{H}) \cap \mathrm{Ker}(\boldsymbol{\Delta}_1) \cap \ldots \cap \mathrm{Ker}(\boldsymbol{\Delta}_N) = \{\boldsymbol{0}_N\}. \tag{116}$$

Note that, for our practical choice for the original image $\boldsymbol{z}$, operator $\boldsymbol{H}$ has full rank and thus (116) holds.

6.1 Majorant mappings and convergence guarantees

The Huber potential terms $(\ell_m)_{1\leq m\leq M}$ satisfy the assumptions from [22, Sec.III] so that we can build the following majorant mapping:

$$(\forall \boldsymbol{x}\in\mathbb{R}^N)\quad \boldsymbol{A}_L(\boldsymbol{x}) = \boldsymbol{H}^\top \mathrm{Diag}\left(\zeta_m([\boldsymbol{H}\boldsymbol{x}]_m])\right)\boldsymbol{H}, \tag{117}$$

with

$$(\forall m=1,\dots,M)(\forall t\in\mathbb{R})\quad \zeta_m(t) = \begin{cases} 1 & \text{if } |t-y_m|\leq p \\ \frac{p}{|t-y_m|} & \text{otherwise.}\end{cases} \tag{118}$$

Function ψ satisfies the properties of [22, Sec.III], allowing us to build a majorant matrix for penalization (115):

$$(\forall \boldsymbol{h}\in\mathbb{R}^N)\quad \boldsymbol{A}_R(\boldsymbol{x}) = \lambda \boldsymbol{\Delta}^\top \mathrm{Diag}(\boldsymbol{\rho}(\boldsymbol{x}))\boldsymbol{\Delta}, \tag{119}$$

with $\boldsymbol{\Delta} = [\boldsymbol{\Delta}_1^\top \mid \dots \mid \boldsymbol{\Delta}_N^\top]^\top \in \mathbb{R}^{2N\times N}$. Moreover,

$$(\forall \boldsymbol{x}\in\mathbb{R}^N)\quad \boldsymbol{\rho}(\boldsymbol{x}) = \begin{bmatrix} \omega(\|\boldsymbol{\Delta}_1\boldsymbol{x}\|)\begin{bmatrix}1\\1\end{bmatrix} \\ \vdots \\ \omega(\|\boldsymbol{\Delta}_N\boldsymbol{x}\|)\begin{bmatrix}1\\1\end{bmatrix}\end{bmatrix} \in \mathbb{R}^{2N}, \tag{120}$$

with $\omega : u\mapsto (1+u^2/\kappa^2)^{-1/2}$. Studying the variations of function ω allows to deduce:

$$(\forall \boldsymbol{x}\in\mathbb{R}^N)\quad \boldsymbol{O}_N \preceq \boldsymbol{A}_R(\boldsymbol{x}) \preceq \lambda\frac{|||\boldsymbol{\Delta}|||^2}{\kappa^2}\boldsymbol{I}_N. \tag{121}$$

In a nutshell, $\boldsymbol{A}_R+\boldsymbol{A}_L$ would constitute a valid majorant mapping for function F. However, it does not necessarily satisfy Assumption 2 since no strictly positive lower-bound is guaranteed for such mapping. We thus hereagain use the corrected mapping:

$$(\forall \boldsymbol{x}\in\mathbb{R}^N)\quad \boldsymbol{A}(\boldsymbol{x}) = \boldsymbol{A}_L(\boldsymbol{x})+\boldsymbol{A}_R(\boldsymbol{x})+\tau\boldsymbol{I}_N, \tag{122}$$

with $\tau>0$. We can thus deduce from (121) and (117) that the mapping (122) satisfies Assumption 2 with:

$$\eta=\tau,\quad \nu=\tau+|||\boldsymbol{H}|||^2+\lambda\frac{|||\boldsymbol{\Delta}|||^2}{\kappa^2}. \tag{123}$$

We focus on the minimization of (112) using the proposed SABRINA scheme for various choices of subspace matrices. We discard the choice $\mathbf{D}_k\equiv\boldsymbol{I}_N$, that appears to be not well suited with such large dimension problem. Instead, we focus on the so-called super-memory gradient subspace family [67], where:

$$(\forall k\in\mathbb{N})\quad \mathbf{D}_k = [-\mathbf{g}_k \,|\, \mathbf{x}_k-\mathbf{x}_{k-1} \,|\, \dots \,|\, \mathbf{x}_{k-M_k+1}-\mathbf{x}_{k-M_k}] \in \mathbb{R}^{N\times M_k}, \tag{124}$$

with the convention $\mathbf{x}_i=\mathbf{0}_N$ for $i<0$, and $M_k\geq 1$ a memory size parameter. The resulting algorithms are denoted SABRINA-SMG-M_k. When $M_k=1$, we retrieve the gradient direction $\mathbf{D}_k=-\mathbf{g}_k$, while for $M_k=2$ we obtain the

memory gradient subspace $\mathbf{D}_k = [-\mathbf{g}_k \,|\, \mathbf{x}_k - \mathbf{x}_{k-1}]$, so that SABRINA-SMG-2 identifies with SABRINA-MG considered in our previous experimental example. Hereagain, if the rank of $\mathbf{D}_k$ gets lower than M_k, columns are removed until satisfying the full column rank assumption. Subspace (124) thus satisfies Assumption 3 for any $M_k \geq 1$. Thus, Theorem 1 and Corollary 1 hold under a moderate gradient noise (see Assumption 4). Assuming than F has isolated stationary points would yield the applicability of Corollary 2, which would guarantee the almost sure convergence of $(\mathbf{x}_k)_{k\in\mathbb{N}}$ to a global minimizer of F. Although it is not possible to show the fulfilment of this technical condition, we did not observe any convergence instability on the sequence $(\mathbf{x}_k)_{k\in\mathbb{N}}$.

6.2 Presentation of the data and settings

The original image $\boldsymbol{z}$ is the satellite image `SanDiego` of size $M = 1024 \times 1024$ pixels. The blur kernel is a non-uniform motion blur with size $N = 21 \times 21$. The noise parameters are $\sigma_1 = 5 \times 10^{-4}$, $\sigma_2 = 200\,\sigma_1$ and $\varrho = 0.1$, so that the signal to noise ratio of the observed image is 13.3 dB. The original image, its degraded version $\mathbf{y}$ and the blur kernel to reconstruct are displayed in Fig. 4.

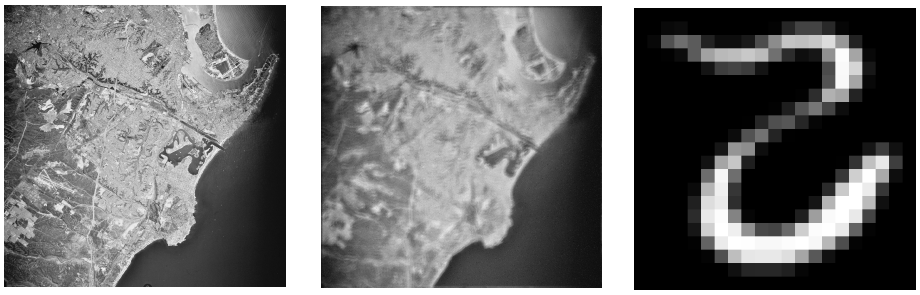

Fig. 4: (Left) Original image $\boldsymbol{z}$; (Middle) Blurred and noisy image $\boldsymbol{y}$; (Right) Original blur kernel $\overline{\boldsymbol{x}}$.

The numerical experiments are performed on the same computer with the same software details as for the example of Section 5. We use the same uniform multiplicative noise (see Sec. 5.2) for the gradient perturbations in our proposed method, so as to satisfy Assumption 4. Once again we set $k = 1/(k+1)^{0.01}$ as the step-size for every $k \in \mathbb{N}$. Finally, the hyperparameters are tuned through gridsearch so as to minimize the relative mean square error (RMSE) on the kernel estimation, to $(p, \lambda, \kappa) = (1, 10, 10)$.

6.3 Calculation of $C_{\max}$

The ratio between bounds (η, ν) involved in (123) allows to compute the allowed tolerance on the gradient uncertainty, following Assumption 4(ii). However, in the

particular problem of blur identification, $|||\boldsymbol{H}|||^2$ may be very large so that $\eta/\nu \ll 1$ and thus $C_{\max} \ll 1$. Typically, in our example, we obtain a theoretical $C_{\max}$ close to 8×10^{-9} which is very constraining in term of gradient noise. Actually, the difficulty lies in the over pessimistic lower bound $\eta = \tau$ in (123). Let us first point out that, according to (117),

$$(\forall k \in \mathbb{N}) \quad (\tau + \min_{1 \leq m \leq M} \zeta_m\left([\boldsymbol{H}\mathbf{x}_k]_m\right) |||\boldsymbol{H}|||^2)\boldsymbol{I}_N \preceq \boldsymbol{A}_k. \tag{125}$$

According to (118), $\min_{1 \leq m \leq M} \zeta_m\left([\boldsymbol{H}\mathbf{x}_k]_m\right) = 1$ as soon as $[\boldsymbol{H}\mathbf{x}_k - \boldsymbol{y}]_m \leq p$ for every $m \in \{1, \ldots, M\}$. This holds for p sufficiently large and/or $\boldsymbol{H}\mathbf{x}_k$ sufficiently close to $\boldsymbol{y}$. We computed the actual values for $\min_{1 \leq m \leq M} \zeta_m\left([\boldsymbol{H}\mathbf{x}_k]_m\right)$ along iterations, in our practical experiment, and observed that this quantity actually goes rapidly to 1 after few iterations. This leads us to consider

$$\widetilde{\eta} = \tau + |||\boldsymbol{H}|||^2 \tag{126}$$

as an empirical lower bound. We denote

$$\widetilde{C}_{\max} = \frac{1}{2}\left(\left(1 + \frac{4\widetilde{\eta}}{\nu}\right)^{\frac{1}{2}} - 1\right), \tag{127}$$

and express the gradient perturbation level C used in the experiment, as a function of $\widetilde{C}_{\max}$. Note that, in the present experiment, $\widetilde{C}_{\max} = 6.18 \times 10^{-1}$, which is closer to the best case bound mentioned in Section 3.3.

6.4 Numerical results

We first compare the performance of SABRINA with classical stochastic algorithms. We also include ADAM method [45], as it shows rather good performance in that example. The gradient perturbation is set to $C = 0.25 \times \widetilde{C}_{\max}$. The methods are compared in terms of RMSE between the current iterate and the sought kernel $\overline{\boldsymbol{x}}$.

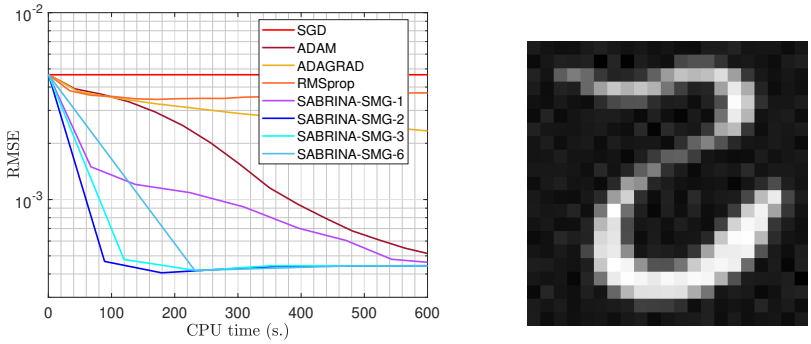

Fig. 5: (Left) Evolution of the RMSE along time for various algorithms ; (Right) Estimated kernel using SABRINA-SMG-2, RMSE $= 4.4 \times 10^{-4}$. Noise amplitude $C = 0.25 \times \widetilde{C}_{\max}$, and starting point $\mathbf{x}_0 = \mathbf{0}_N$.

Fig. 5(left) shows that SABRINA-SMG-2 is the fastest of the algorithms to reach convergence. The other choices of memory size, for the super-memory gradient subspace, appear less competitive, which is in accordance with the observations from [22,25]. The RMSE of the reconstructed kernel, displayed in Fig. 5(right), is equal to 4.4×10^{-4} for an estimated computational time of 600 s.

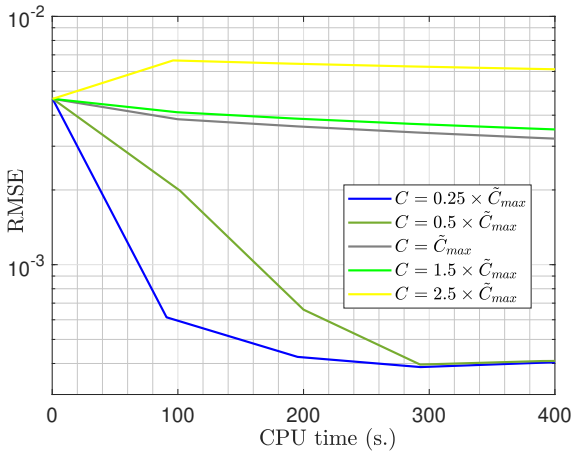

Fig. 6: Evolution of the RMSE along time for various noise amplitudes affecting the gradient term in SABRINA-SMG-2.

For the setting of Fig. 5(left), the value of C actually exceeds the maximal numerical tolerance $C_{\max}$, but is chosen lower than $\widetilde{C}_{\max}$ which appears sufficient in practice to reach convergence for all SABRINA variants tested here. In Fig. 6, we now present the evolution of the RMSE along time for SABRINA-SMG-2, when its gradient term is affected by various levels of noise C. One can notice that our corrected bound $\widetilde{C}_{\max}$ clearly maps with the delineation of two regimes for the convergence of SABRINA-SMG-2. As soon as C is sufficiently low compared to $\widetilde{C}_{\max}$, convergence is fast. On the contrary, a too high C seems to compromise the behaviour of the method, as expected and divergence can even be observed for large C.

7 Conclusion

Our work provides new insights into the stability of MM schemes suffering from stochastic noise perturbations on their gradient evaluation. New asymptotical results and convergence rate analysis are demonstrated under reasonably mild assumptions, and in the challenging scenario of a non necessarily convex cost function. Two numerical experiments in the fields of machine learning and image processing illustrate the high relevancy of the considered MM schemes compared to several classical competitors both regarding their speed of convergence and their robustness to noise. In particular, the experimental results emphasize the impressive performance of MM algorithm associated to a memory gradient subspace,

already assessed in our previous works [22, 25]. It is remarkable to notice that, for such subspace choice, our contribution can be understood as providing novel theoretical guarantees on a stochastic non-linear conjugate gradient method with MM-based formula for stepsize and conjugacy parameters. One avenue for future work would be to extend our convergence rate analysis to a larger class of function by alleviating the strong convexity condition.

Declarations

Funding: This research work received funding support from the European Research Council Starting Grant MAJORIS ERC-2019-STG-850925.
Competing interests: The authors have no relevant financial or non-financial interests to disclose.
Data availability statement: The datasets generated during and/or analysed during the current study are available from the corresponding author on reasonable request.

Acknowledgement

The authors thank Jean-Christophe Pesquet (Univ. Paris Saclay, France) and Matthieu Terris (Heriot-Watt University, UK) for initial motivations and thoughtful discussions.

References

1. P.-A. Absil and K. A. Gallivan. Accelerated line-search and trust-region methods. *SIAM Journal on Numerical Analysis*, 47(2):997–1018, 2009.
2. Ö. D. Akyildiz, E. Chouzenoux, V. Elvira, and J. Míguez. A probabilistic incremental proximal gradient method. *IEEE Signal Processing Letters*, 26(8):1257–1261, 2019.
3. M. Allain, J. Idier, and Y. Goussard. On global and local convergence of half-quadratic algorithms. *IEEE Transactions on Image Processing*, 15(5):1130–1142, 2006.
4. Y. F. Atchadé, G. Fort, and E. Moulines. On perturbed proximal gradient algorithms. *Journal on Machine Learning Research*, 18:1–33, 2017.
5. T. Bell, J. Xu, and S. Zhang. Method for out-of-focus camera calibration. *Applied Optics*, 55(9):3246–2352, 2016.
6. D. P. Bertsekas. *Nonlinear Programming*. Athena Scientific, Belmont, Massachusetts, 3rd edition, 2016.
7. D. P. Bertsekas and J. N. Tsitsiklis. Gradient convergence in gradient methods with errors. *SIAM Journal on Optimization*, 10(3):627–642, 2000.
8. R. Bhatia. *Matrix Analysis*, volume 169. Springer Science & Business Media, 2013.
9. J. Bolte and E. Pauwels. Majorization-minimization procedures and convergence of SQP methods for semi-algebraic and tame programs. *Mathematics of Operations Research*, 41(2):442–465, 2016.
10. J.-F. Bonnans, J. C. Gilbert, C. Lemaréchal, and C. A. Sagastizábal. *Numerical Optimization: Theoretical and Practical Aspects*. Springer Science & Business Media, 2006.
11. A. Bordes, L. Bottou, and P. Gallinari. SGD-QN: Careful quasi-Newton stochastic gradient descent. *Journal on Machine Learning Research*, 10:1737–1754, Jul. 2009.
12. L. Bottou. Large-scale machine learning with stochastic gradient descent. In Y. Lechevallier and G. Saporta, editors, *Proceedings of COMPSTAT 2010*, pages 177–186, Heidelberg, 2010. Physica-Verlag HD.
13. L. Bottou, F. E. Curtis, and J. Nocedal. Optimization methods for large-scale machine learning. *Siam Review*, 60(2):223–311, 2018.

14. G. Bouchard. Efficient bounds for the softmax function and applications to approximate inference in hybrid models. In *Proceedings of the Neural Information Processing Systems (NIPS 2008)*, volume 31, Vancouver, B.C., Canada, 8-11 Dec. 2008.
15. L. M. Briceño-Arias, G. Chierchia, E. Chouzenoux, and J.-C. Pesquet. A random block-coordinate Douglas–Rachford splitting method with low computational complexity for binary logistic regression. *Computational Optimization and Applications*, 72(3):707–726, 2019.
16. R. H. Byrd, S. L. Hansen, J. Nocedal, and Y. Singer. A stochastic quasi-Newton method for large-scale optimization. *SIAM Journal on Optimization*, 26(2):1008–1031, 2016.
17. S. Cadoni, E. Chouzenoux, J.-C. Pesquet, and C. Chaux. A block parallel majorize-minimize memory gradient algorithm. In *Proceedings of the 23rd IEEE International Conference on Image Processing (ICIP 2016)*, pages 3194–3198, Phoenix, AZ, 25-28 Sep. 2016.
18. E. J. Candes, M. B. Wakin, and S. P. Boyd. Enhancing sparsity by reweighted l1 minimization. *Journal of Fourier Analysis and Applications*, 14(5-6):877–905, 2008.
19. C. Castera, J. Bolte, C. Fevotte, and E. Pauwels. An inertial Newton algorithm for deep learning. Technical report, 2019. https://arxiv.org/abs/1905.12278.
20. M. Chalvidal and E. Chouzenoux. Block distributed 3MG algorithm and its application to 3D image restoration. In *2020 IEEE International Conference on Image Processing (ICIP)*, pages 938–942, Abu Dhabi, United Arab Emirates (virtual), 25-28 Oct. 2020.
21. C.-C. Chang and C.-J. Lin. Libsvm: A library for support vector machines. *ACM Transactions on Intelligent Systems and Technology*, 2(3):1–27, 2011. Software available at *http://www.csie.ntu.edu.tw/ cjlin/libsvm*.
22. E. Chouzenoux, J. Idier, and S. Moussaoui. A majorize–minimize strategy for subspace optimization applied to image restoration. *IEEE Transactions on Image Processing*, 20(6):1517–1528, 2010.
23. E. Chouzenoux, A. Jezierska, J.-C. Pesquet, and H. Talbot. A majorize-minimize subspace approach for $\ell_2 - \ell_0$ image regularization. *SIAM Journal on Imaging Sciences*, 6(1):563–591, 2013.
24. E. Chouzenoux and J.-C. Pesquet. Convergence rate analysis of the majorize-minimize subspace algorithm. *IEEE Signal Processing Letters*, 23(9):1284–1288, Sep. 2016.
25. E. Chouzenoux and J.-C. Pesquet. A stochastic majorize-minimize subspace algorithm for online penalized least squares estimation. *IEEE Transactions on Signal Processing*, 65(18):4770–4783, 2017.
26. P. L. Combettes and J.-C. Pesquet. Stochastic approximations and perturbations in forward-backward splitting for monotone operators. *Pure Applied Functional Analysis*, 1(1):13–37, Jan. 2016.
27. B. Delyon, M. Lavielle, and E. Moulines. Convergence of a stochastic approximation version of the EM algorithm. *The Annals of Statistics*, 27(1):94–128, 1999.
28. A. Dieuleveut, A. Durmus, and F. Bach. Bridging the gap between constant step size stochastic gradient descent and markov chains. *arXiv preprint arXiv:1707.06386*, 2017.
29. J. Duchi, E. Hazan, and Y. Singer. Adaptive subgradient methods for online learning and stochastic optimization. *Journal of Machine Learning Research*, 12(7), 2011.
30. V. Dudar, G. Chierchia, E. Chouzenoux, J.-C. Pesquet, and V. Semenov. A two-stage subspace trust region approach for deep neural network training. In *Proceedings of the 25th European Signal Processing Conference (EUSIPCO 2017)*, Kos Island, Greece, 28 Aug.-2 Sep. 2017.
31. M. Duflo. *Random Iterative Models*. Stochastic Modelling and Applied Probability. Springer Berlin, Heidelberg, 1st edition, 2013.
32. V. Elvira and E. Chouzenoux. Optimized population Monte Carlo. *IEEE Transactions on Signal Processing*, 2022. (to appear).
33. J. M. Ermoliev and Z. V. Nekrylova. The method of stochastic gradients and its application. In *Seminar: Theory of Optimal Solutions. No. 1 (Russian)*, pages 24–47. Akad. Nauk Ukrain. SSR, Kiev, 1967.
34. B. Fehrman, B. Gess, and A. Jentzen. Convergence rates for the stochastic gradient descent method for non-convex objective functions. *Journal of Machine Learning Research*, 21, 2020.
35. J. Fernandez-Bes, V. Elvira, and S. Van Vaerenbergh. A probabilistic least-mean-squares filter. In *2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 2199–2203, Brisbane, Australia, 19-24 Apr. 2015.

36. J. Fest and E. Chouzenoux. Stochastic majorize-minimize subspace algorithm with application to binary classification. In *Proceedings of the 29th European Signal Processing Conference (EUSIPCO 2021)*, Dublin, Ireland (virtual), 23-27 Aug. 2021.
37. A. Florescu, E. Chouzenoux, J.-C. Pesquet, P. Ciuciu, and S. Ciochina. A majorize-minimize memory gradient method for complex-valued inverse problems. *Signal Processing*, 103:285–295, 2014.
38. S. Gadat. Stochastic optimization algorithms, non asymptotic and asymptotic behaviour. *Lecture notes, University of Toulouse*, 2017.
39. S. Gadat and I. Gavra. Asymptotic study of stochastic adaptive algorithm in non-convex landscape. Technical report, 2021. https://arxiv.org/abs/2012.05640.
40. D. Geman and C. Yang. Nonlinear image recovery with half-quadratic regularization. *IEEE Transactions on Image Processing*, 4(7):932–946, 1995.
41. M. Gharbi, E. Chouzenoux, J.-C. Pesquet, and L. Duval. GPU-based implementations of MM algorithms. Application to spectroscopy signal restoration. In *Proceedings of the 29th European Signal Processing Conference (EUSIPCO 2021)*, Dublin, Ireland (virtual), 23-27 Aug. 2021.
42. I. Gitman, H. Lang, P. Zhang, and L. Xiao. Understanding the role of momentum in stochastic gradient methods. In *Advances in Neural Information Processing Systems*, pages 9633–9643, 2019.
43. Y. Huang, E. Chouzenoux, and V. Elvira. Probabilistic modeling and inference for sequential space-varying blur identification. *IEEE Transactions on Computational Imaging*, 7:531–546, 2021.
44. M. W. Jacobson and J. A. Fessler. An expanded theoretical treatment of iteration-dependent Majorize-Minimize algorithms. *IEEE Transations on Image Processing*, 16(10):2411–2422, Oct. 2007.
45. D. P. Kingma and J. Ba. Adam: A method for stochastic optimization. Technical report, 2014. https://arxiv.org/abs/1412.6980.
46. J. Konečnỳ, J. Liu, P. Richtárik, and M. Takáč. Mini-batch semi-stochastic gradient descent in the proximal setting. *IEEE Journal of Selected Topics in Signal Processing*, 10(2):242–255, 2015.
47. C. Li, C. Chen, D. Carlson, and L. Carin. Preconditioned stochastic gradient langevin dynamics for deep neural networks. Technical report, 2015. https://arxiv.org/abs/1512.07666.
48. D. C. Liu and J. Nocedal. On the limited memory bfgs method for large scale optimization. *Mathematical programming*, 45(1):503–528, 1989.
49. N. Loizou and P. Richtárik. Momentum and stochastic momentum for stochastic gradient, newton, proximal point and subspace descent methods. *Computational Optimization and Applications*, 77(3):653–710, 2020.
50. J. Mairal. Stochastic majorization-minimization algorithms for large-scale optimization. Technical report, 2013. https://arxiv.org/abs/1306.4650.
51. Y. Marnissi, E. Chouzenoux, A. Benazza-Benyahia, and J.-C. Pesquet. Majorize-minimize adapted Metropolis-Hastings algorithm. *IEEE Transactions on Signal Processing*, (68):2356–2369, Mar. 2020.
52. P.-A. Meyer. *Martingales and Stochastic Integrals I*. Springer Berlin, Heidelberg, 1st edition, 2006.
53. A. Miele and J. Cantrell. Study on a memory gradient method for the minimization of functions. *Journal of Optimization Theory and Applications*, 3(6):459–470, 1969.
54. E. Moulines and F. Bach. Non-asymptotic analysis of stochastic approximation algorithms for machine learning. *Advances in neural information processing systems*, 24, 2011.
55. Y. E. Nesterov. A method for solving the convex programming problem with convergence rate $O(1/k^2)$. In *Dokl. akad. nauk Sssr*, volume 269, pages 543–547, 1983.
56. J. Nocedal and S. Wright. *Numerical optimization*. Springer Science & Business Media, 2006.
57. A. N. Ostrowski. *Solutions of Equations in Euclidean and Banach Spaces*. Academic Press, 1973.
58. M. Pereyra, P. Schniter, E. Chouzenoux, J.-C. Pesquet, J.-Y. Tourneret, A. O. Hero, and S. McLaughlin. A survey of stochastic simulation and optimization methods in signal processing. *IEEE Journal of Selected Topics in Signal Processing*, 10(2):224–241, 2015.
59. A. Repetti and Y. Wiaux. A forward-backward algorithm for reweighted procedures: Application to radio-astronomical imaging. In *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP 2020)*, pages 1434–1438, 4-8 May 2020.

60. E. Richardson, R. Herskovitz, B. Ginsburg, and M. Zibulevsky. Seboost-boosting stochastic learning using subspace optimization techniques. Technical report, 2016. https://arxiv.org/abs/1609.00629.
61. H. Robbins and S. Monro. A stochastic approximation method. *The Annals of Mathematical Statistics*, pages 400–407, 1951.
62. H. Robbins and D. Siegmund. *A convergence theorem for non negative almost supermartingales and some applications*, pages 233–257. Academic Press, 1971.
63. M. C. Robini and Y. Zhu. Generic half-quadratic optimization for image reconstruction. *SIAM Journal on Imaging Sciences*, 8(3):1752–1797, 2015.
64. L. Rosasco, S. Villa, and B. Vu. Convergence of stochastic proximal gradient algorithm. *Applied Mathematics and Optimization*, 82:891–917, 2020.
65. S. Ruder. An overview of gradient descent optimization algorithms. Technical report, 2016. https://arxiv.org/abs/1609.04747.
66. M. Sghaier, E. Chouzenoux, J.-C. Pesquet, and S. Muller. A novel task-based reconstruction approach for digital breast tomosynthesis. *Medical Image Analysis*, 77:102341, 2022.
67. Z.-J. Shi and J. Shen. Convergence of supermemory gradient method. *Journal of Applied Mathematics and Computing*, 24(1):367–376, 2007.
68. P. Sonneveld. CGS: A fast Lanczos-type solver for nonsymmetric linear systems. *SIAM Journal on Scientific Statistical Computing*, 10(1):36–52, Jan. 1989.
69. Y. Sun, P. Babu, and D. P. Palomar. Majorization-minimization algorithms in signal processing, communications, and machine learning. *IEEE Transactions on Signal Processing*, 65(3):794–816, 2016.
70. I. Sutskever, J. Martens, G. Dahl, and G. Hinton. On the importance of initialization and momentum in deep learning. In *Proceedings of the International conference on machine learning (ICML 2013)*, pages 1139–1147, Atlanta, USA, 16-21 June 2013.
71. T. Tieleman and G. Hinton. RMSProp: Divide the gradient by a running average of its recent magnitude. *COURSERA: Neural Networks for Machine Learning*, 6(5), 2012.
72. Y.-X. Yuan. Subspace techniques for nonlinear optimization. In *Some Topics in Industrial and Applied Mathematics*, pages 206–218. World Scientific, 2007.
73. Z. Zhang, J. T. Kwok, and D.-Y. Yeung. Surrogate maximization/minimization algorithms and extensions. *Machine Learning*, 69:1–33, 2007.
74. M. Zibulevsky. SESOP-TN: Combining sequential subspace optimization with truncated newton method. Technical report, 2008. http://www.optimization-online.org/DB_FILE/2008/09/2098.pdf.