Stochastic Majorize-Minimize Subspace Algorithm with Application to Binary Classification

Jean-Baptiste Fest and Émilie Chouzenoux
Université Paris-Saclay, Inria, CentraleSupélec, Centre de Vision Numérique, firstname.name@centralesupelec.fr

***Abstract*—In a learning context, data distribution are usually unknown. Observation models are also sometimes complex. In an inverse problem setup, these facts often lead to the minimization of a loss function with uncertain analytic expression. Consequently, its gradient cannot be evaluated in an exact manner. These issues have has promoted the development of so-called stochastic optimization methods, which are able to cope with stochastic errors in the gradient term. A natural strategy is to start from a deterministic optimization approach as a baseline, and to incorporate a stabilization procedure (e.g., decreasing stepsize, averaging) that yields improved robustness to stochastic errors. In the context of large-scale, differentiable optimization, an important class of methods relies on the principle of majorization-minimization (MM). MM algorithms are becoming increasingly popular in signal/image processing [18], [36] and machine learning [27], [34], [38]. MM approaches are fast, stable, require limited manual settings, and are often preferred by practitioners in application domains such as medical imaging [16] and telecommunications [29]. The present work introduces novel theoretical convergence guarantees for MM algorithms when approximate gradient terms are employed, generalizing some recent work [11], [27] to a wider class of functions and algorithms. We illustrate our theoretical results with a binary classification problem.**

***Index Terms*—Stochastic optimization, convergence analysis, Majorization-Minimization, subspace acceleration, binary logistic regression.**

I. Introduction

A common strategy to find a relevant solution to supervised learning problems and inverse problems relies on the minimization of a loss function. This function integrates knowledge about the available data/models, and some prior information on the sought parameters. This yields the generic problem

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

where $F : \mathbb{R}^N \mapsto \mathbb{R}$ is differentiable on $\mathbb{R}^N$. A major challenge arises when the analytical properties of F, and, more importantly, of its gradient ∇F, cannot be precisely known. For instance, in the context of supervised learning, $\boldsymbol{x}$ corresponds to a set of parameters to be learned. Moreover, F is interpreted as a random expectation, making an exact evaluation of F and ∇F either impossible or too computationally intensive. Consequently, only a stochastic approximation can used instead [2]. In the context of inverse problems, stochastic errors mainly arise from two sources: (i) sophisticated acquisition models, requiring for instance solving a PDE for evaluating the fidelity to data [25], and (ii) on-the-fly (i.e., online) data processing resulting from joint acquisition and reconstruction process, imposed in some specific application contexts (e.g., medical imaging [24]). Here again, the evaluation of F and ∇F is approximate, which jeopardizes the stability of the retained optimization solver. As a response to this challenge, it is necessary to design large scale optimization tools whose robustness to stochastic errors is guaranteed. Although extending deterministic optimization methods to a probabilistic context does not in principle affect the structure of their original version, the analysis for asymptotical guarantees in the stochastic context requires novel theoretical analysis making use of specific tools from stochastic approximation theory [32].

A large amount of studies in stochastic optimization literature have been dedicated to the famous gradient descent algorithm [3], [17], [31] and its various first order variations such as NAG [28]. One can also mention ADAM, ADAGRAD, and RMSprop methods [14], [26], [37], widely adopted in the field of deep neural network. The account for non-differentiable terms, through proximal-based approaches, has also been investigated in [1], [12], [33]. Another avenue of works adopt the key concept of majorization-minimization (MM). MM algorithms rely on successive majorizing approximations of the function in order to produce a sequence of iterates that will converge to a solution of the problem, under suitable assumptions. The loss function F is thus substituted, by a sequence of surrogates with better properties (typically, strongly convex quadratic functions), ensuring sounded theoretical stability and fast practical convergence to the resulting scheme. The theoretical robustness of MM approaches to stochastic errors has been studied in [11], [13], [27]. However, the aforementioned works are, up to our knowledge, over specific in the MM scheme and/or the loss function class they studied. For instance, [13] focused on the expectation-minimization framework, while [27] requires averaging procedures over consecutive majorizing approximation and iterates. Finally, [11] deals with online recursive least-squares problems enjoying specific recursive forms for the gradient and majorizing approximations.

In this work, we propose a more versatile formulation of the problem, when F is a non necessarily convex, smooth function. We follow an approach where we build, at each iteration, a stochastic quadratic surrogate involving a randomly

This work is funded by the European Research Council Starting Grant MAJORIS ERC-2019-STG-850925

perturbed evaluation of the current gradient. This yields an inexact stochastic MM scheme, in which we also allow the possibility for a subspace acceleration [8], [30], in the spirit of momentum-based approaches [15], [21], [23], [26], [28]. We establish almost sure convergence results for our approach, under mild assumptions on F. We illustrate the validity of the proposed scheme and its great performance in a problem of large scale binary logistic regression.

The paper is organized as follows. Section II introduces the MM framework and the considered inexact version for it. Section III presents our main contribution, that is the convergence analysis for this algorithm. Numerical experiments are depicted in Section IV. Finally, Section V concludes the paper.

II. Proposed formulation

A. Deterministic MM approach

The MM algorithm is an iterative process which solves (1) by generating a sequence $(\boldsymbol{x}_k)_{k\in\mathbb{N}} \in \mathbb{R}^N$ of iterates. Passing from a iterate $\boldsymbol{x}_k$ to the next one $\boldsymbol{x}_{k+1}$ for a given $k \in \mathbb{N}$ is made by minimizing h, a tractable majorant approximation of F. Typical choice is to resort to a quadratic approximation for F, reading similarly as a second-order Taylor expansion of it, at a given point $\boldsymbol{y} \in \mathbb{R}^N$:

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

Here above, $\|\cdot\|^2_{\boldsymbol{A}(\boldsymbol{y})} = \langle \cdot | \boldsymbol{A}(\boldsymbol{y}) \cdot \rangle$, and $\boldsymbol{A} : \boldsymbol{y} \mapsto A(\boldsymbol{y})$ is a function returning a symmetric positive definite matrix such that $(\forall \boldsymbol{x}, \boldsymbol{y} \in \mathbb{R}^N) \quad h(\boldsymbol{x}, \boldsymbol{y}) \geq F(\boldsymbol{x})$. This mapping, also called the majorant metric mapping, completely describes h and thus the quality of the majorizing approximation of F. Consequently, the generic MM update reads:

$$(\forall k \in \mathbb{N}) \quad \boldsymbol{x}_{k+1} = \boldsymbol{x}_k - \boldsymbol{A}_k^{-1} \nabla F(\boldsymbol{x}_k), \quad (3)$$

denoting $\boldsymbol{A}_k := \boldsymbol{A}(\boldsymbol{x}_k)$. The above MM scheme yields, by construction, the monotonic decrease of $(F(\boldsymbol{x}_k))_{k\in\mathbb{N}}$. Useful strategies to build the majorant metric can be found in [9], [36], [38], for a large class of problems arising in applications from supervised learning, telecommunications, and image restoration, to name a few. In the latter case, the algorithm in Eq. (3) is also known as half-quadratic method [22], highly popular in the 90s.

B. Subspace acceleration

As can be seen, the minimization of the surrogate at each current iterate requires the inversion of an $N \times N$ operator. However, the commonly very large number of parameters encountered in practical situations (e.g., in 3D image restoration, $N \geq 10^9$) could make such an operation hazardous. One major key of improvement introduced by [8], and later assessed in the survey paper [36], consists in integrating so-called "subspace acceleration" [30] in the update (3). Mathematically, this amounts to defining

$$(\forall k \in \mathbb{N}) \qquad \boldsymbol{x}_{k+1} = \boldsymbol{x}_k + \boldsymbol{D}_k \boldsymbol{u}_k, \quad (4)$$

with

$$(\forall k \in \mathbb{N}) \quad \boldsymbol{u}_k \in \underset{\boldsymbol{u}\in\mathbb{R}^{M_k}}{\arg\min}\, h(\boldsymbol{x}_k + \boldsymbol{D}_k \boldsymbol{u}, \boldsymbol{x}_k). \quad (5)$$

Here above, $\boldsymbol{D}_k \in \mathbb{R}^{N\times M_k}$, $N \geq M_k \geq 1$ is a new degree of freedom in the approach. Columns of matrix $\boldsymbol{D}_k$ contain a set of search directions to explore and in order to define the new iterate $\boldsymbol{x}_{k+1}$. Case of $M_k = N$ and $\boldsymbol{D}_k = \boldsymbol{I}_N$ the identity matrix of $\mathbb{R}^N$ obviously leads back to Algorithm (3). More interesting choices are listed in [8, Tab.1]. In particular, $\boldsymbol{D}_k = [-\nabla F(\boldsymbol{x}_k) \quad | \quad \boldsymbol{x}_k - \boldsymbol{x}_{k-1}]$ leads to the so-called MM Memory Gradient (3MG) algorithm [9], [10] whose great performances have been assessed in [9], [19]. It is worth noting that the quadratic structure of h makes a solution $\boldsymbol{u}_k$ to (5) easy to be determined as:

$$(\forall k \in \mathbb{N}) \quad \boldsymbol{u}_k = -\left(\boldsymbol{D}_k^\top \boldsymbol{A}_k \boldsymbol{D}_k\right)^\dagger \boldsymbol{D}_k^\top \nabla F(\boldsymbol{x}_k), \quad (6)$$

with $\dagger$ being the pseudo-inverse operation. The convergence properties of (4)-(6) are discussed in [9], [10].

C. Stochastic subspace MM scheme

In this paper, we focus on the introduction of a perturbation on F (and its gradient). More precisely, we will consider that, at each iteration $k \in \mathbb{N}$, one only has access to:

$$\boldsymbol{g}_k = \nabla F(\boldsymbol{x}_k) + \boldsymbol{\epsilon}_k, \quad (7)$$

with $(\boldsymbol{\epsilon}_k)_{k\in\mathbb{N}} \in \mathbb{R}^N$ modeling a stochastic noise process with zero mean and bounded variance, which will be specified in the next section. In order to write the stochastic version of (4)-(6), we need to introduce the concept of inexact majorant function. Let $k \in \mathbb{N}$. We can evaluate the stochastic quadratic function $\widehat{h}_k$ defined by:

$$\widehat{h}_k : \boldsymbol{u} \in \mathbb{R}^N \mapsto F(\boldsymbol{x}_k) + \boldsymbol{g}_k^\top (\boldsymbol{u} - \boldsymbol{x}_k) + \frac{1}{2}\|\boldsymbol{u} - \boldsymbol{x}_k\|^2_{\boldsymbol{A}_k}. \quad (8)$$

Similarly to its deterministic counterpart, the proposed stochastic MM subspace scheme will consist in finding the next iterate according to the available $\widehat{h}_k$, within the subspace spanned by the columns of $\boldsymbol{D}_k$. We will assume, without loss of generality, that $\boldsymbol{D}_k$ has full column rank. Moreover, for better stability of the iterates, we introduce a positive step-size sequence $(\gamma_k)_{k\in\mathbb{N}}$. This leads to our novel stochastic MM subspace algorithm:

$$(\forall k \in \mathbb{N}) \quad \boldsymbol{x}_{k+1} = \boldsymbol{x}_k + \gamma_k \boldsymbol{D}_k \widehat{\boldsymbol{u}}_k, \quad (9)$$

where

$$(\forall k \in \mathbb{N}) \quad \widehat{\boldsymbol{u}}_k = -\left(\boldsymbol{D}_k^\top \boldsymbol{A}_k \boldsymbol{D}_k\right)^{-1} \boldsymbol{D}_k^\top \boldsymbol{g}_k. \quad (10)$$

D. Link with preconditioned gradient method

One may combine (9)-(10) to reach the more compact structure:

$$(\forall k \in \mathbb{N}) \quad \boldsymbol{x}_{k+1} = \boldsymbol{x}_k - \gamma_k \boldsymbol{B}_k \boldsymbol{g}_k, \quad (11)$$

with $\boldsymbol{B}_k = \boldsymbol{D}_k \left(\boldsymbol{D}_k^\top \boldsymbol{A}_k \boldsymbol{D}_k\right)^{-1} \boldsymbol{D}_k^\top$. The main interest of the latter formulation comes from its similarity with a preconditioned stochastic gradient structure [4], [6]. The symmetric

matrix $\boldsymbol{B}_k \in \mathbb{R}^{N\times N}$ gathers both the information given by the majorant matrix $\boldsymbol{A}_k$ and the subspace matrix $\boldsymbol{D}_k$. To a certain extent, understanding the behaviour of $\boldsymbol{B}_k$ should allow to control the convergence of sequence $(\boldsymbol{x}_k)_{k\in\mathbb{N}}$. The theoretical challenges to tackle are twofold: (i) $\boldsymbol{B}_k$ is random and non necessarily full rank, (ii) F is non necessarily convex. Up to our knowledge, the convergence of the scheme (11) has never been studied under such context.

III. Asymptotical analysis

Let us introduce $(\Omega, \mathcal{F}, P)$ a probability space provided with the filtration $(\mathcal{F}_k)_{k\in\mathbb{N}}$ where $\mathcal{F}_0 = \{\Omega, \emptyset\}$ and for all $k \geq 1$, $\mathcal{F}_k = \sigma(\boldsymbol{\epsilon}_0, \boldsymbol{x}_1, ..., \boldsymbol{\epsilon}_{k-1}, \boldsymbol{x}_k)$ is the sub-sigma algebra generated by $\{\boldsymbol{\epsilon}_0, \boldsymbol{x}_1, ..., \boldsymbol{\epsilon}_{k-1}, \boldsymbol{x}_k\}$, gathering all of the information available at time k. For all $k \in \mathbb{N}$, we denote by $\mathbb{E}(.|\mathcal{F}_k)$, the conditional expectancy operator relative to $\mathcal{F}_k$. A property will be said to be satisfied *almost surely* (a.s) if it holds on a probability-one set of $\mathcal{F}$.

A. Assumptions

We enumerate here the necessary technical assumptions for the establishment of our main convergence theorem. A discussion around these assumptions will be provided in Section III-C.

Assumption 1: F is coercive, differentiable on $\mathbb{R}^N$, with a bounded gradient along the iterates, i.e., there exists $G > 0$ such that, for every $k \in \mathbb{N}$, $\|\nabla F(\boldsymbol{x}_k)\| \leq G$ a.s.

Assumption 2: There exists $(\eta, \nu) > 0$ such that, for every $k \in \mathbb{N}$, $\eta \boldsymbol{I}_N \preceq \boldsymbol{A}_k \preceq \nu \boldsymbol{I}_N$ a.s.

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

(i) $\operatorname{rank}(\boldsymbol{D}_k) = M_k$ a.s.

(ii) $\boldsymbol{g}_k \in \operatorname{Ker}(\boldsymbol{D}_k^\top)^\perp$ a.s.

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

(i) $(\forall k \in \mathbb{N})$ $\mathbb{E}(\boldsymbol{\epsilon}_k|\mathcal{F}_k) = 0$ a.s.

(ii) There exists $C \in [0, C_{\max}[$ with $C_{\max} = \frac{1}{2}((1+\frac{4\eta}{\nu})^{\frac{1}{2}} - 1)$ such that:

$$(\forall k \in \mathbb{N}) \quad \mathbb{E}\left(\|\boldsymbol{\epsilon}_k\|^2|\mathcal{F}_k\right) \leq C^2\|\nabla F(\boldsymbol{x}_k)\|^2 \quad \text{a.s.} \quad (12)$$

Assumption 5: $\sum_{k=0}^{\infty} \gamma_k = +\infty, \quad \sum_{k=0}^{+\infty} \gamma_k^2 < +\infty.$

B. Convergence results

We start by an intermediary result, regarding the behaviour of $(\boldsymbol{B}_k)_{k\in\mathbb{N}}$.

Lemma 1: Under Assumptions 2 and 3(i), (10) is a.s well defined and, for every $k \in \mathbb{N}$,

$$\begin{cases} \boldsymbol{O} \preceq \boldsymbol{B}_k \preceq \frac{1}{\eta}\boldsymbol{I}_N \quad \text{a.s.} \\ (\forall \boldsymbol{x} \in \operatorname{Ker}(\boldsymbol{D}_k^\top)^\perp) \quad \boldsymbol{x}^\top \boldsymbol{B}_k \boldsymbol{x} \geq \frac{1}{\nu}\|\boldsymbol{x}\|^2 \quad \text{a.s.} \end{cases}$$

Let us now state our main convergence result:

Theorem 1: Assume that $(\boldsymbol{x}_k)_{k\in\mathbb{N}}$ satisfies scheme (9)-(10) and Assumptions 1-5 are verified. Then the following holds :

(i) The sequence $(F(\boldsymbol{x}_k))_{k\in\mathbb{N}}$ converges a.s to an a.s finite random variable.

(ii) $\liminf_{k\to+\infty} \|\nabla F(\boldsymbol{x}_k)\| = 0$ a.s.

The proof, that we skipped by lack of space, is made of two main steps. First, we rely on Lemma 1 to obtain a specific stochastic relationship between two consecutive iterates. Second, we make use of the Robbins-Siegmund's lemma of [32] which leads to Theorem 1(i) and to a Zoutendijk type condition. The latter allows us to deduce Theorem 1(ii).

Theorem 1 promotes an interesting general behaviour (i.e a sub-sequence criterion) which is nonetheless not sufficient to guarantee convergence of $(\boldsymbol{x}_k)_{k\in\mathbb{N}}$ to a minimizer or even to a stationary point of F. To this aim, we need additional topological assumptions on F and on its set of stationary points. Let us denote $\operatorname{zer}\nabla F$ such a set. Moreover, for every $\boldsymbol{v} \in \mathbb{R}^N$, we introduce $\operatorname{lev}_{=v} F := \left\{\boldsymbol{x} \in \mathbb{R}^N \;\middle|\; F(\boldsymbol{x}) = \boldsymbol{v}\right\}$, the level set of F relative to $\boldsymbol{v}$. Then, we can derive the following convergence theorem.

Theorem 2: Assume that assumptions of Theorem 1 hold on $\mathbb{R}^N$. Then:

(i) If all level set of F are finite, $(\boldsymbol{x}_k)_{k\in\mathbb{N}}$ converges a.s to a stationary point of F.

(ii) If F is convex on $\mathbb{R}^N$ and all of its stationary points are isolated i.e $\forall v \in \mathbb{R} \quad \operatorname{Card}(\operatorname{lev}_{=v} F \cap \operatorname{zer}\nabla F) < +\infty$, then $(\boldsymbol{x}_k)_{k\in\mathbb{N}}$ converge a.s to a minimizer of F.

Remark that a strongly-convex F sees the sequence $(\boldsymbol{x}_k)_{k\in\mathbb{N}}$ converging almost surely to its unique minimizer as a direct corollary of Theorem 2(ii).

C. Discussion about the assumptions

Assumption 1 and 5 are rather standard in the analysis of stochastic gradient-based methods. Note that Assumption 1 is less restrictive than the Lipschitz continuity, assumed for instance in [23], [27]. Assumptions 2 and 3 are typically required in the convergence analysis of quadratic MM subspace methods [8]. Assumption 3(ii) holds for instance when $-\boldsymbol{g}_k$ belongs to the range of $\boldsymbol{D}_k$. Assumption 4(i) corresponds to a standard hypothesis when it comes to study stochastic process optimization. Assumption 4(ii), reminiscent from [23], relates to the second order moment of $\boldsymbol{\epsilon}_k$ and expresses that the uncertainty affecting the gradient should remain moderate, in regards with the norm of the (true) gradient. The upper bound $C_{\max}$ is as large as the condition number η/ν of the majorant metric sequence is. The theoretical maximal value corresponding to $\frac{\sqrt{5}-1}{2} \simeq 6,1\times10^{-1}$ is obtained when $\eta = \nu$. This can happen when $\boldsymbol{A}_k \equiv \eta\boldsymbol{I}_N$ with some $\eta > 0$ (for e.g., η can be taken as the Lipschitz constant of ∇F, when it exists). However, such a choice goes back to a basic gradient descent scheme, which can have limited practical convergence speed. In contrast, $C_{\max} \sim \eta/\nu$ for $\eta/\nu \to 0^+$, which shows that poorly conditioned majorant matrices would impose a high demanding bound on the gradient error as a balance. A compromise is thus necessary, between the sophistication of the majorant metric and its condition number.

D. Link to existing works

The theoretical result closer to our Theorem 1 is the one obtained in [21], [23]. These works address an algorithm similar to the famous ADAM [26], that can actually be viewed

as a particular case of ours, when no MM metric is used (i.e., $\boldsymbol{A}_k \equiv \boldsymbol{I}_N$) and a 3MG-like subspace is employed associated to manually tuned stepsize and momentum weight. As already mentioned, the method from [27], also shares connection with our algorithm, since it includes an MM strategy, though not combined with any subspace acceleration. A convergence result, similar to our Theorem 1, is obtained in [27], with a slightly less tractable subsequence criterion (see [27, Prop.3.3]). Finally, the approach from [4], [6] can be understood as a particular choice for the subspace, related to quasi-Newton approximation of F, but without the use of MM metric/stepsize. The authors show an ℓ_1 behaviour for the sequence $(F(\boldsymbol{x}_k))_{k\in\mathbb{N}}$, however under more restrictive assumptions (e.g., convexity of F). The use of the level-set hypothesis (see Theorem 2(i)) is reminiscent from [20], [21]. Finally, it is worth noting that our study, and in particular Theorem 2(ii), generalizes the conclusions that were obtained in [11] for a specific class of loss function, gradient error and majorant approximation.

IV. Application to supervised classification

A. Problem statement

We consider the problem of supervised binary classification through regularized logistic regression. Starting from a training dataset made of m feature vectors $(\boldsymbol{v}_i)_{1\leq i\leq m} \in \mathbb{R}^N$, and their associated labels $(y_i)_{1\leq i\leq m} \in \{-1,1\}$, we learn the parameters $\boldsymbol{x} \in \mathbb{R}^N$ of a linear classifier by minimizing the following penalized empirical risk:

$$(\forall \boldsymbol{x} \in \mathbb{R}^N) \quad F(\boldsymbol{x}) = \frac{1}{m}\sum_{i=1}^{m} \log\left(1+\exp\left(-[\boldsymbol{L}\boldsymbol{x}]_i\right)\right)+\frac{\mu}{2}\|\boldsymbol{x}\|^2. \tag{13}$$

Here above, $\boldsymbol{L} = \mathrm{Diag}\{(y_i)_{1\leq i\leq m}\}[\boldsymbol{v}_1,...,\boldsymbol{v}_m]^\top \in \mathbb{R}^{m\times N}$ and μ is a positive regularization parameter aiming at limiting overfitting effects. We can obtain majorant mapping $\boldsymbol{A}(\cdot)$ following [5]:

$$(\forall \boldsymbol{x} \in \mathbb{R}^N) \quad \boldsymbol{A}(\boldsymbol{x}) = \boldsymbol{L}^\top \mathrm{Diag}\{(\omega([\boldsymbol{L}\boldsymbol{x}]_i)_{1\leq i\leq m}\}\boldsymbol{L} + \mu \boldsymbol{I}_N, \tag{14}$$

where $\omega : x \mapsto \frac{1}{x}\left(\frac{1}{1+\exp(-x)} - \frac{1}{2}\right)$, extended by continuity so that $\omega(0) = \frac{1}{4}$. Let us note that such mapping satisfies Assumption 2 for any $\mu > 0$, since we have:

$$(\forall \boldsymbol{x} \in \mathbb{R}^N) \quad \mu \boldsymbol{I}_N \preceq \boldsymbol{A}(\boldsymbol{x}) \preceq \left(\mu + \frac{1}{4m}\|\boldsymbol{L}\|^2\right)\boldsymbol{I}_N. \tag{15}$$

Constant $\|\boldsymbol{L}\|^2/4m$ relates to the conditioning of the training dataset. The smaller it is, the higher the permissible noise value (see Ass. 4(ii)).

B. Numerical settings

The algorithms are implemented in Matlab 2020a and run on a desktop having an Intel Core i7 3.2 GHz pro with 16 GB of RAM. We consider the perturbed MM scheme (9)-(10), for two choices for the subspace $\boldsymbol{D}_k$ both satisfying Assumption 3, namely $\boldsymbol{D}_k = \boldsymbol{I}_N$, and $\boldsymbol{D}_k = [-\boldsymbol{g}_k \mid \boldsymbol{x}_k - \boldsymbol{x}_{k-1}]$, yielding the so-called SMM and S3MG algorithms, respectively. For the former, the inversion of $\boldsymbol{A}_k$ is performed using the solver [35]. The decreasing stepsize sequence $\gamma_k = 1/(k+1)^{0.51}$ is used, so as to verify Assumption 5. Comparisons are made with several state-of-the-art stochastic algorithms, namely SGD [3], ADAM [26], ADAGRAD [14] and RMSprop [37]. The tuning for their parameters (e.g., learning rate, momentum weight) was made empirically so as to reach best convergence profiles. Function F is strongly convex as far as $\mu > 0$, so the convergence of the sequence generated by SMM and S3MG to the unique solution of (1) is ensured, as long as Assumption 4 holds. In our experiments, multiplicative noise following a uniform law centered in 1 (so that Ass. 4(i) holds) is added on each component of the gradient. Several noise amplitudes will be tested, satisfying (12), with $C \geq 0$ satisfying or not the range constraint imposed in Assumption 4 (ii) (see hereafter for more details).

We use `rcv1` and `a8a` datasets from LIBSVM library [7], whose properties are summarized in table below. Parameter μ was manually set to get a good accuracy for the classifier on the test set, when training without noise perturbation on the gradient loss. This leads to an accuracy of 0.92 and 0.82 on test set, for `rcv1` and `a8a`, respectively.

Train Size m	Test Size	Features N	$\|\boldsymbol{L}\|^2/(4m)$	μ
20242	677399	47236	$5,5\times10^{-3}$	1
9865	22696	122	1,6	10^{-1}

C. Experimental results

Fig. 1 illustrates the performance of the different methods, in terms of gradient norm evolution along time. In this case, we set $C \simeq 0.9 \times C_{\max}$ so that Ass. 4(ii) holds, and thus the convergence of S3MG and SMM is ensured. It is remarkable that both largely outperform their competitors. Moreover, one can see the advantage brought by the subspace acceleration in both examples. Finally, let us emphasize that the implementation of the MM methods did not require any tedious manual tuning. Fig. 2 shows the evolution of the gradient norm, when using S3MG with different noise levels, on the `rcv1` example. Faster convergence is reached for lower noise amplitudes, as expected. Moreover, one can see that S3MG starts showing oscillating behaviour when $C \geq C_{\max}$. This shows our theoretical bound $C_{\max}$ is valid and not over pessimistic for guaranteeing practical stability of the method.

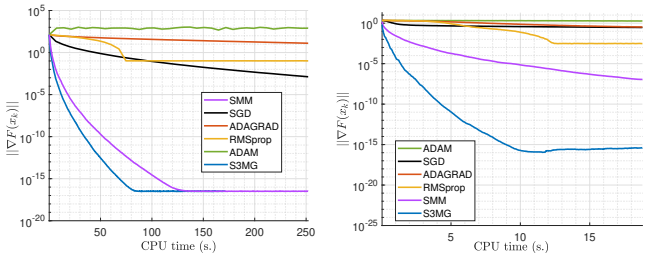

Fig. 1. Evolution of the gradient norm along time for various algorithms, on dataset `rcv1` (left) and `a8a` (right). Noise amplitude $C \simeq 0.9 \times C_{\max}$.

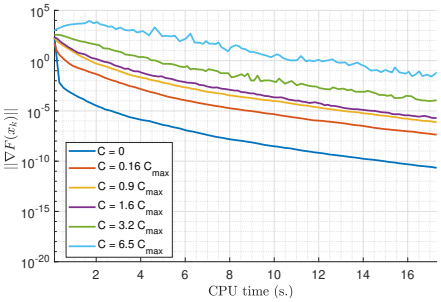

Fig. 2. `rcv1`: Evolution of the gradient norm along time for various noise amplitudes using S3MG algorithm.

V. Conclusion

This work sheds some new light on the stability of MM quadratic schemes in the presence of a stochastic error on the gradient evaluation. New asymptotic results are obtained under mild assumptions. Our numerical application aims at illustrating the great performance of MM schemes in a supervised learning context, both in terms of convergence speed and stability, when compared to several competitors. Future work will be dedicated to convergence rate analysis.

References

[1] Y. F. Atchadé, G. Fort, and E. Moulines. On perturbed proximal gradient algorithms. *Journal on Machine Learning Research*, 18:1–33, 2017.

[2] F. Bach and E. Moulines. Non-asymptotic analysis of stochastic approximation algorithms for machine learning. In *Proceedings of the Annual Conference on Neural Information Processing Systems (NIPS 2011)*, pages x–x+8, Granada, Spain, Dec. 12 - 17 2011.

[3] D. P. Bertsekas and J. N. Tsitsiklis. Gradient convergence in gradient methods with errors. *SIAM Journal on Optimization*, 10(3):627–642, 2000.

[4] 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.

[5] 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, 2008.

[6] 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.

[7] 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*.

[8] 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.

[9] 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.

[10] 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.

[11] 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.

[12] 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.

[13] 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.

[14] J. Duchi, E. Hazan, and Y. Singer. Adaptive subgradient methods for online learning and stochastic optimization. *Journal of Machine Learning Research*, 12(7), 2011.

[15] 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.

[16] H. Erdogan and J. A. Fessler. Monotonic algorithms for transmission tomography. *IEEE Transations on Medical Imaging*, 18(9):801–814, Sept. 1999.

[17] 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.

[18] M. Figueiredo, J. Bioucas-Dias, and R. Nowak. Majorization-minimization algorithms for wavelet-based image restoration. *IEEE Transactions on Image Processing*, 16(12):2980–2991, Dec. 2007.

[19] 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.

[20] S. Gadat. Stochastic optimization algorithms, non asymptotic and asymptotic behaviour. *Lecture notes, University of Toulouse*, 2017.

[21] S. Gadat and I. Gavrat. Asymptotic study of stochastic adaptive algorithm in non-convex landscape. Technical report, 2021. https://arxiv.org/abs/2012.05640.

[22] D. Geman and C. Yang. Nonlinear image recovery with half-quadratic regularization. *IEEE Transactions on Image Processing*, 4(7):932–946, 1995.

[23] 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.

[24] L. E. Gueddari, E. Chouzenoux, A. Vignaud, J.-C. Pesquet, and P. Ciuciu. Online MR image reconstruction for compressed sensing acquisition in T2* imaging. In *Proceedings of SPIE 11138, Wavelets and Sparsity XVIII*, volume 1113819, Sep. 2019.

[25] M. Huska, D. Lazzaro, S. Morigi, A. Samore, and G. Scrivanti. Spatially-adaptive variational reconstructions for linear inverse electrical impedance tomography. *Journal on Scientific Computing*, 84(46), 2020.

[26] D. P. Kingma and J. Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

[27] J. Mairal. Stochastic majorization-minimization algorithms for large-scale optimization. *arXiv preprint arXiv:1306.4650*, 2013.

[28] 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.

[29] D. Ramirez, I. Santamaria, L. L. Scharf, and S. Van Vaerenbergh. Multi-channel factor analysis with common and unique factors. *IEEE Transactions on Signal Processing*, 68:113–126, 2020.

[30] E. Richardson, R. Herskovitz, B. Ginsburg, and M. Zibulevsky. Seboost-boosting stochastic learning using subspace optimization techniques. *arXiv preprint arXiv:1609.00629*, 2016.

[31] H. Robbins and S. Monro. A stochastic approximation method. *The Annals of Mathematical Statistics*, pages 400–407, 1951.

[32] H. Robbins and D. Siegmund. A convergence theorem for non negative almost supermartingales and some applications. In *Optimizing methods in statistics*, pages 233–257. Elsevier, 1971.

[33] V. S. Rosasco, L. and B. Vu. Convergence of stochastic proximal gradient algorithm. *Applied Mathematics and Optimization*, 82:891–917, 2020.

[34] V. Singhal and A. Majumdar. Majorization minimization technique for optimally solving deep dictionary learning. *Neural Processing Letters*, 47:799–814, Jun. 2018.

[35] P. Sonneveld. CGS: A fast Lanczos-type solver for nonsymmetric linear systems. *SIAM Journal on Scientific Statistical Computing*, 10(1):36–52, Jan. 1989.

[36] 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.

[37] 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.

[38] Z. Zhang, J. T. Kwok, and D.-Y. Yeung. Surrogate maximization/minimization algorithms and extensions. *Machine Learning*, 69:1–33, 2007.