Fast Stochastic AdaGrad for Nonconvex Bound-Constrained Optimization using Second-Order Information

Stefania Bellavia[*], Serge Gratton[†], Benedetta Morini[‡], Philippe L. Toint[§]

30 I 2026

Abstract

ADAGB2, a generalization of the AdaGrad algorithm for stochastic optimization is introduced, which handles bound-constrained problems and uses second-order information when available. It is shown that, given $\delta \in (0, 1)$ and $\epsilon \in (0, 1]$, the ADAGB2 algorithm needs at most $\mathcal{O}(\epsilon^{-2})$ iterations to ensure an ϵ-approximate first-order critical point of the bound-constrained problem with probability at least $1 - \delta$ under a new "directional" assumption on the gradient oracle and its conditional root-mean-square error. Should this latter condition fail, it is also shown that the optimality level of iterates is bounded above by this average. The relation between the approximate and true classical projected-gradient-based optimality measures for bound constrained problems is also investigated, and it is shown that merely assuming unbiased gradient oracles may be insufficient to ensure convergence in $\mathcal{O}(\epsilon^{-2})$ iterations.

Keywords: AdaGrad, stochastic nonlinear optimization, objective-function-free optimization (OFFO), complexity, second-order information, stochastic projected gradients, bound constraints.

1 Introduction

First-order optimization algorithms have been widely used in the contexts on online learning and deep neural network training and their convergence properties on nonconvex problems have been investigated by several authors (see [9] for a survey). Among them, AdaGrad [19], despite not being always as efficient as others in practice on nonconvex problems [43], enjoys a special position from the theoretical point of view because of its solid and extensive convergence analysis.

When the objective function's gradient is contaminated by noise (for instance caused by sampling) a probabilistic point of view on the algorithm's convergence theory is desirable. This has been investigated for AdaGrad applied to nonconvex functions by a number of authors, as shown in Table 1, along with the relevant important assumptions and results. We discuss this rich body of theory (and the content of this table) in more detail in Section 1.1.

These proposals are however limited, from the theoretician's point of view, in three respects.

[*]Dipartimento di Ingegneria Industriale, Università degli Studi di Firenze, Firenze, Italia. Member of the INdAM Research Group GNCS. Email: stefania.bellavia@unifi.it.

[†]Université de Toulouse, INP, IRIT, Toulouse, France. Work partially supported by the Artificial and Natural Intelligence Toulouse Institute (ANITI), French "Investing for the Future - PIA3" program under the Grant agreement ANR-19-PI3A-0004. Email: serge.gratton@enseeiht.fr.

[‡]Dipartimento di Ingegneria Industriale, Università degli Studi di Firenze, Firenze, Italia. Member of the INdAM Research Group GNCS. Email: benedetta.morini@unifi.it.

[§]Namur Center for Complex Systems (naXys), University of Namur, Namur, Belgium.Work partially supported by the Artificial and Natural Intelligence Toulouse Institute (ANITI). Email: philippe.toint@unamur.be.

Paper		Smooth	Gradients bound / bias	Noise type	2nd-order usage	Bound constr.	Conv. type	Conv. rate
Li-Orabona	[31]	L^*	no / no	Sub-Gauss. + $\sigma \searrow 0$	no	no	$\mathbb{E}[.]$/w.h.p.	$\mathcal{O}_\ell(\epsilon^{-2})$
Ward et al.	[42]	L	yes / no	Bounded	no	no	w.h.p.	$\mathcal{O}_\ell(\epsilon^{-4})$
Défossez et al.	[18]	L	yes / no	Unrestricted	no	no	$\mathbb{E}[.]$	$\mathcal{O}_\ell(\epsilon^{-4})$
Gratton et al.	[23]	L	yes / no	Unrestricted	no	no	$\mathbb{E}[.]$	$\mathcal{O}_\ell(\epsilon^{-4})$
Kavis et al.	[27]	L	yes / no	Sub-Gauss. + $\sigma \searrow 0$	no	no	w.h.p	$\mathcal{O}_\ell(\epsilon^{-2})$
Faw et al.	[21]	L	no / no	Affine	no	no	w.h.p.	$\mathcal{O}_\ell(\epsilon^{-4})$
Wang et al.	[38]	(L_0, L_1^*)	no / no	Affine + $\sigma \searrow 0$	no	no	$\mathbb{E}[.]$/w.h.p.	$\mathcal{O}_\ell(\epsilon^{-2})$
Liu et al.	[32]	L	no / no	Sub-Gauss. + $\sigma \searrow 0$	no	no	w.h.p.	$\mathcal{O}(\epsilon^{-2})$
Attia-Koren	[3]	L	no / no	Affine-* + $\sigma \searrow 0$	no	no	w.h.p.	$\mathcal{O}_\ell(\epsilon^{-2})$
Faw et al.	[20]	(L_0, L_1^*)	no / no	Affine	no	no	w.h.p.	$\mathcal{O}_\ell(\epsilon^{-4})$
Hong-Lin	[25]	L	no / no	Affine-* + $\sigma \searrow 0$	no	no	w.h.p.	$\mathcal{O}_\ell(\epsilon^{-2})$
Hong-Lin	[25]	(L_0, L_1^*)	no / no	Affine-*	no	no	w.h.p.	$\mathcal{O}_\ell(\epsilon^{-4})$
Jiang et al.	[26]	L	no / no	Bounded	no	no	$\mathbb{E}[.]$	$\mathcal{O}_\ell(\epsilon^{-2})$
Alacaoglu et al.	[2, 1]	L	yes/yes	Bounded	no	yes	$\mathbb{E}[.]$	$\mathcal{O}_M(\epsilon^{-4})$
This paper		L	no / no	new1	yes	yes	$\mathbb{E}[.]$/w.h.p	$\mathcal{O}(\epsilon^{-2})$
This paper		L	no / yes	new1+new2	yes	yes	$\mathbb{E}[.]$/w.h.p	$\mathcal{O}(\epsilon^{-2})$

Table 1: Convergence theories for the stochastic AdaGrad algorithm and their characteristics in the nonconvex setting. (In the third column, L^* means that the Lipschitz constant must be known, the last column reports a bound (in order) for the considered algorithm to achieve ϵ-criticality, the subscript ℓ indicating the presence of an additional logarithmic factor in ϵ, the subscript M indicates that the convergence is studied in terms of the norm of the gradient of the Moreau envelope, see Section 1.1 for more detail on the other columns).

- The first is that the probabilistic analysis is restricted to the case where second-order information is ignored when available. The algorithm remains strictly first-order, with a step always aligned (possibly component-wise) with the approximate steepest-descent direction (no trust region is used). By contrast, other first-order methods have been "augmented" to use step-sizes along the first-order direction taking second-order information into account (see [5, 7, 15, 17, 36, 39] and the references therein).

- The second is the capability of solving constrained problems that arise when *a priori* information on the problem at hand is available, often in the form of constraints, of which bounds on the variables are the most common. Such problems arise, for instance in fracture mechanics [28, 29], inverse problems and PDE-constrained optimization under uncertainty [12]. Moreover, bounds are frequently applied to machine learning models to avoid overfitting [30] or to reflect real-world limits and maintain physical interpretability [33]. Admittedly, these can be taken care of using penalty terms in the loss/objective function, as is often done for Physically Informed Neural Networks (PINNs) (see [11, 41] for instance), but this introduces new hyper-parameters needing calibration and does not ensure that constraints are strictly satisfied. Moreover, a strong penalization of the constraints degrades the problem's conditioning, possibly causing slow convergence, especially if first-order methods are used.

  Focusing on Adagrad for constrained problems, we are aware of only [1, 2] that provide the stochastic analysis of a projected version of Adagrad-Norm. More generally, stochastic gradient-based methods for constrained problems have been proposed, e.g., in [1, 14, 16, 22, 40] and references there-in, but these techniques differ considerably from the AdaGrad-like gradient algorithms considered here.

- The third shortcoming is that, to the best of the authors' knowledge, bias in the gradient oracle has been considered only for AdaGrad-Norm in [1, 2], in contrast with other stochastic optimization techniques (see [6, 8] for instance).

The challenge addressed in this paper is thus threefold. We first consider a probabilistic theory of how second-order steps can be accommodated in a stochastic AdaGrad-like[1] optimization method. We also show how this can be done in the context of bound-constrained problems, and, finally, propose a theory which does not assume that gradient oracles are unbiased.

More specifically, we consider the problem

$$\min_{x \in \mathcal{F}} f(x) \quad \text{where} \quad \mathcal{F} = \{x \in \mathbb{R}^n \mid l_i \le x_i \le u_i\} \tag{1}$$

where f is a smooth (possibly nonconvex) function from an open set containing the feasible region $\mathcal{F} \subseteq \mathbb{R}^n$ into $\mathbb{R}$, and where l and u are vectors specifying the lower and upper bounds on the variables, respectively (infinite entries in l and/or u are allowed). We also assume that $\nabla_x^1 f(x)$ cannot be computed exactly but is approximated by a random oracle. As a consequence, the algorithm we are about to describe generates a random process, where, for a given iterate x_k, the oracle computes the gradient oracle $g(x_k, \xi)$ where ξ is a random variable (whose distribution may depend on x_k), with probability space $(\Omega, \mathcal{T}_\Omega, \mathbb{P})$. For brevity, we will denote $g_k \stackrel{\rm def}{=} g(x_k, \xi)$ and we also define $G_k = \nabla_x^1 f(x_k)$ and $H_k = \nabla_x^2 f(x_k)$. Conditioned to knowing $g_0, \ldots, g_{k-1}$, the expectation will be denoted by the symbol $\mathbb{E}_k[\cdot]$, the probability of an event will be denoted as $P_k(\cdots)$ and $\mathbb{1}(\cdot)$ will denote the indicator function of the event. The symbol $\|\cdot\|$ denotes the 2-norm.

1.1 Related works

The convergence of AdaGrad in the nonconvex setting has been studied by a number of authors assuming the use of an unbiased stochastic oracle of the true gradient with a variety of assumptions on its noise. In [31], a version of AdaGrad with delayed step-sizes was introduced. The almost sure asymptotic convergence of the gradients to zero was proved for this variant assuming that the noise has bounded support, i.e.,

$$\mathbb{E}_k[\|g_k - G_k\|] \le \kappa_{\rm bounded} \qquad \kappa_{\rm bounded} > 0,$$

at any iteration k. Assuming instead that, at each iteration k, the sub-Gaussian noise condition

$$\mathbb{E}_k\left[e^{\|g_k - G_k\|^2/\sigma^2}\right] \le e,$$

holds, Li and Orabona also showed that AdaGrad is noise adaptive, in the sense that, given $\epsilon > 0$, the iteration complexity interpolates, with high probability (denoted by w.h.p. in Table 1), between $\mathcal{O}_\ell(\epsilon^{-4})$ to $\mathcal{O}_\ell(\epsilon^{-2})$ depending on the magnitude of σ, where $\mathcal{O}_\ell(\cdot)$ means that the order is up to a logarithmic term. Unfortunately, this analysis requires *a priori* knowledge of the Lipschitz constant of the gradient in setting the step-size. A similar result was proved in [27] but without assuming this knowledge. Ward, Wu and Bottou in [42] analyzed the convergence of AdaGrad-Norm assuming uniformly bounded gradients and bounded variance of the gradient oracle at each iteration k, that is

$$\mathbb{E}_k\left[\|g_k - G_k\|^2\right] \le \sigma^2.$$

They showed that the average of the squared norm of the gradients produced by AdaGrad-Norm converges with high probability at the rate $\mathcal{O}(1/\sqrt{k})$, which implies that the algorithm needs at most $\mathcal{O}_\ell(\epsilon^{-4})$ iterations to achieve $\mathbb{E}[\|G_k\|] \le \epsilon$. Under weaker assumptions, i.e., without assuming boundedness of the variance of the gradient's oracles, these results have been extended (now in expectation form, noted by $\mathbb{E}[\cdot]$ in Table 1) to the component-wise version of AdaGrad in [18] and to an extended class of methods comprising AdaGrad in [23].

A different stream of research analyzed AdaGrad as an optimally-tuned adaptive stochastic gradient descent without assuming bounded gradients but under the affine bound on the variance at each iteration given by

$$\mathbb{E}_k\left[\|g_k - G_k\|^2\right] \le \kappa_{1,\rm affine} + \kappa_{2,\rm affine}\|G_k\|^2, \tag{2}$$

[1]That is an algorithm that reduces to AdaGrad when second-order information is not used.

with $\kappa_{1,\text{affine}} \geq 0$ and $\kappa_{2,\text{affine}} \geq 0$. Under these assumptions, Faw et al. [21] have shown that AdaGrad-Norm exhibits an iteration complexity of the order of $\mathcal{O}_\ell(\epsilon^{-4})$, with high probability.

The noise adaptivity of AdaGrad under the stronger affine condition

$$\|g_k - G_k\|^2 \overset{as}{\leq} \kappa_{1,\text{affine}} + \kappa_{2,\text{affine}}\|G_k\|^2 \tag{3}$$

(denoted Affine-* in Table 1) has been proved in [3]. The same complexity bound for AdaGrad-Norm is obtained in [38] under the affine condition (2).

Very recently, Jiang, Maladkar and Mokhtari in [26] performed the convergence analysis in 1-norm and proved $\mathcal{O}_\ell(\epsilon^{-2})$ iteration complexity, assuming coordinate-wise bounded variance and coordinate-wise Lipschitz continuity of the gradient.

Finally Hong and Lin revisited the convergence of AdaGrad in the recent paper [25], assuming a relaxed version of the condition (3) given by

$$\|g_k - G_k\|^2 \overset{as}{\leq} \kappa_{1,\text{affine}} + \kappa_{2,\text{affine}}\|G_k\|^2 + \kappa_{3,\text{affine}}(f(x_k) - f_{\text{low}}), \tag{4}$$

where f_{low} is such that $f(x) \geq f_{\text{low}}\ \forall x \in \mathbb{R}^n$. They prove that the iteration complexity interpolates between $\mathcal{O}_\ell(\epsilon^{-4})$ to $\mathcal{O}_\ell(\epsilon^{-2})$ depending on the magnitude of $\kappa_{1,\text{affine}}$ and $\kappa_{3,\text{affine}}$.

Projected versions of AdaGrad-Norm have been proposed in [2, 1]. In these latter papers biased stochastic oracles of the gradient are allowed. The paper [2] elaborates on [16] and proves the $\mathcal{O}(\epsilon^{-4})$ iteration complexity employing as optimality measure the norm of the gradient of Moreau envelope. The analysis works with mini-batch size of 1 and assuming bounded stochastic gradients. This analysis has been further extended under general non-i.i.d. data sampling assumption in [1].

All the previous mentioned results have been obtained assuming the Lipschitz continuity of the gradient. The papers [20, 25, 38] provide results for the class of (L_0, L_1)-smooth functions[2] and assume that the constant L_1 is known. In [20] the results in [21] have been extended to this latter class of functions and it has been proved that the method has a complexity of the order of $\mathcal{O}_\ell(\epsilon^{-4})$, assuming (2) with $\kappa_{2,\text{affine}} < 1$. Similar results are given in [25] for AdaGrad with momentum under the Assumption (4). Stronger results are obtained in [38] for AdaGrad-Norm assuming (2) where it is proved that the iteration complexity interpolates between $\mathcal{O}_\ell(\epsilon^{-4})$ to $\mathcal{O}_\ell(\epsilon^{-2})$ depending on the magnitude of $\kappa_{1,\text{affine}}$.

To conclude this brief survey, we also note that Liu et al. [32] have obtained an $\mathcal{O}(\epsilon^{-2})$ convergence result for the objective-function gap produced by the stochastic AdaGrad-Norm when applied to γ-quasar convex L-smooth functions with a sub-Weibull assumption on the gradient error. This condition subsumes the sub-Gaussian case, but the result does not apply to general L-smooth nonconvex functions.

1.2 Summary of contributions

In view of all contributions discussed above, we summarize our contributions as follows.

1. We propose an extension of AdaGrad which is capable of using (possibly very approximate) second-order information whenever available.

2. This algorithm is also suitable for problems involving bounds on the variables.

3. We analyze the convergence and probabilistic complexity of this algorithm, showing that it solves the approximate minimization problem with optimal complexity under a new directional condition on the gradient error, but without assuming bounded or unbiased gradient oracles nor knowledge of the problem's Lipschitz constant.

4. Taking a more general perspective, we discuss the relation between the classical optimality measure for bound-constrained problems induced by the projected gradient in the stochastic

[2] A function is (L_0, L_1)-smooth if there exist a constant $L_0 \geq 0$ and $L_1 \geq 0$ and $\delta > 0$ such that for all $x, y \in \mathbb{R}^n$ with $\|x - y\| \leq \delta$, one has that $\|\nabla_x^1 f(x) - \nabla_x^1 f(y)\| \leq (L_0 + L_1\|\nabla_x^1 f(y)\|)\|x - y\|$.

case, and have shown that, in general, the unbiased nature of the gradient oracle is not sufficient to ensure convergence on the true problem with the optimal rate, even if such a convergence occurs for the approximate one.

5. We furthermore show that our optimal complexity result for the approximated problem extends to the true problem (i.e. using exact gradients) if another condition on the gradient noise holds or if the gradient oracle is unbiased and the problem is unconstrained. We also describe the behaviour of the algorithm if none of these conditions hold.

6. As far as the authors are aware, this is the first convergence-rate result for the stochastic AdaGrad applied to bound constrained problems and allowing to use second-order information.

Section 2 presents the new algorithm and discusses its features, while convergence analysis for the approximate problem is described in Section 3. Section 4 discusses the relation between approximate and true optimality measures in the bound-constrained case and its application to the new algorithm. A brief conclusion and some perspectives are finally outlined in Section 5.

2 The algorithm

Our proposed algorithm, called ADAGB2, is presented on this page. Its inputs consist in a starting point $x_{\rm ini}$ and values l and u for the lower and upper bounds on the variables.

Algorithm 2.1: ADAGB2$(x_{\rm ini}, l, u)$

Step 0: Initialization: The constants $\varsigma, \tau \in (0,1]$ and $\kappa_s \geq 1$ are given.
Set $x_0 = P_{\mathcal{F}}(x_{\rm ini})$, $k = 0$ and $w_{-1,i} = \sqrt{\varsigma}$ for $i \in \{1,\ldots,n\}$.

Step 1: First-order step: Compute $g_k = g(x_k, \xi)$ a random approximation of G_k, set

$$d_k \stackrel{\rm def}{=} P_{\mathcal{F}}(x_k - g_k) - x_k, \tag{5}$$

$$w_{k,i} = \sqrt{w_{k-1,i}^2 + d_{k,i}^2} \;\text{ and }\; \Delta_{k,i} = \frac{|d_{k,i}|}{w_{k,i}} \quad \text{for} \quad i \in \{1,\ldots,n\}. \tag{6}$$

$$\mathcal{B}_k = \{x \in \mathbb{R}^n \mid |x_i - x_{k,i}| \leq \Delta_{k,i} \;\text{ for }\; i \in \{1,\ldots,n\}\} \tag{7}$$

and

$$s_k^L = P_{\mathcal{F}\cap\mathcal{B}_k}(x_k - g_k) - x_k. \tag{8}$$

Step 2: Second-order step: Choose B_k a symmetric approximation of H_k and compute

$$s_k^Q = \gamma_k s_k^L \;\text{ where }\; \gamma_k = \begin{cases} \min\left[1, \dfrac{-g_k^T s_k^L}{(s_k^L)^T B_k s_k^L}\right] & \text{if} \quad (s_k^L)^T B_k s_k^L > 0 \\ 1 & \text{otherwise.} \end{cases} \tag{9}$$

Then select s_k such that, for all $i \in \{1,\ldots,n\}$,

$$x_k + s_k \in \mathcal{F}, \quad |s_{k,i}| \leq \kappa_s \Delta_{k,i} \quad \text{and} \quad g_k^T s_k + \tfrac{1}{2} s_k^T B s_k \leq \tau \left(g_k^T s_k^Q + \frac{1}{2}(s_k^Q)^T B s_k^Q \right). \tag{10}$$

Step 3: Loop: Set $x_{k+1} = x_k + s_k$, increment k by one and go to Step 1.

The statement of the algorithm suggests the following comments.

1. We first note that the mechanism of the algorithm ensures that all iterates remain feasible, that is $x_k \in \mathcal{F}$ for all $k \geq 0$.

2. The projections $P_{\mathcal{F}}$ and $P_{\mathcal{F}\cap\mathcal{B}_k}$ occurring in (5) and (8) are extremely cheap to compute component-wise, as, for any vector $y \in \mathbb{R}^n$ and $i \in \{1,\ldots,n\}$,

$$[P_{\mathcal{F}}(y)]_i = \max[l_i, \min[y_i, u_i]] \tag{11}$$

and

$$[P_{\mathcal{F}\cap\mathcal{B}_k}(y)]_i = \max[l_i, x_{k,i} - \Delta_{k,i}, \min[y_i, x_{k,i} + \Delta_{k,i}, u_i]]. \tag{12}$$

Note that (8) can be rewritten as

$$s_k^L = \mathcal{P}_{\mathcal{F}\cap\mathcal{B}_k}(x_k - g_k) - x_k = \mathcal{P}_{\mathcal{B}_k}(\mathcal{P}_{\mathcal{F}}(x_k - g_k)) - x_k = \mathcal{P}_{\mathcal{B}_k}(x_k + d_k) - x_k.$$

3. The uniform bound specified by AS.3 below is the only restriction made on B_k (beyond symmetry). This allows for a wide range of deterministic or random approximations, such as (sampled) Barzilai-Borwein, safeguarded (limited-memory) quasi-Newton, random sketching or finite-difference approximations of $(s_k^L)^T B_k s_k^L$ using one evaluation of the gradient oracle. Obviously, exact Hessians may also be used when available.

4. The shifted quadratic model of the objective function given by $g_k^T s_k + \frac{1}{2} s_k^T B s_k$ is minimized by $x_k + s_k^Q$ along the intersection of the first-order direction s_k^L with the trust region $\mathcal{B}_k$. In the vocabulary of trust-region methods, it can therefore be interpreted as the "Cauchy point" at iteration k (see [13, Sections 6.3 and 12.2.1]). Since $s_k = s_k^Q$ satisfies (10), an improved second-order step s_k minimizing the quadratic model $g_k^T s + \frac{1}{2} s^T B s$ beyond $x_k + s_k^Q$ is possible and often beneficial. This is standard and well-tested practice in truncated Newton or second-order trust-region or adaptive-regularization algorithms. The idea is to apply a conjugate-gradient or Lanczos method, which performs successive minimizations of the model in nested Krylov subspaces, until a desired accuracy threshold is reached (see for instance [37, 35] or [13, Section 7.5] for detailed descriptions and algorithms). Note that these techniques do not require computing a possibly approximate Hessian matrix, but only require its product with vectors. However, this remains optional in ADAGB2 (as is the evaluation of a nonzero B_k and the computation of s_k^Q itself) and the choice $B_k = 0$ is possible when access to second-order information is too expensive. Should this choice be made, (9) gives that $s_k^Q = s_k^L$ and the choice $s_k = s_k^L$ is always acceptable for (10), in which case ADAGB2 is a purely first-order algorithm.

5. If the choice $B_k = 0$ is made at all iterations and the problem is unconstrained in that $l_i = -\infty$ and $u_i = +\infty$ for all i, then $d_k = -g_k$ and thus ADAGB2 is nothing but the standard AdaGrad algorithm. In short,

$$\text{ADAGB2} = \text{AdaGrad} + \text{2nd order} + \text{bound constraints},$$

the last two items being of course optional. As a consequence, the stochastic complexity theory described below applies to AdaGrad without any modification.

3 Convergence analysis

The convergence theory we are about to describe is based on the following assumptions.

AS.1: *The function f is twice continuously differentiable and the feasible region $\mathcal{F}$ is not empty.*

AS.2: *There exists a constant $L \geq 0$ such that for all $x, y \in \mathbb{R}^n$*

$$\|\nabla_x^1 f(x) - \nabla_x^1 f(y)\| \leq L\|x - y\|.$$

We stress that, although the existence of L is assumed, the knowledge of its value is *not* needed to run the algorithm.

AS.3: *There exists a constant $\kappa_B \geq 1$ such that $\|B_k\| \leq \kappa_B$* for all $k \geq 0$.

AS.4: *The objective function is bounded below on the feasible domain, that is there exists a constant $f_{\rm low} < f(x_0)$ such that $f(x) \geq f_{\rm low}$ for every $x \in \mathcal{F}$.*

AS.5: *There exists a constant $\kappa_{Gg} > 0$ such that*

$$\mathbb{E}_k\left[|(G_k - g_k)^T s_k|\right] \leq \kappa_{Gg}^2 \mathbb{E}_k\left[\|s_k\|^2\right]$$

for all $k \geq 0$.

This condition (which we have called "new1" in Table 1) can be interpreted as a "root mean square" condition along the direction s_k (see the comments after Theorem 4.2 below). As far as the authors are aware, this condition is necessary if a Cauchy point is introduced to take second-order information into account because γ_k then becomes a random variable.

Assumption AS.5 requires increasing accuracy when the iterates converge. This is needed in our analysis to estimate the conditional expectation of $G_k^T s_k$ when the step is no longer a scaled multiple of the negative gradient, but can be quite general provided it produces a sufficient decrease in the quadratic model decrease. Thus AS.5 is admittedly stronger than the typical bounded variance condition assumed in the simpler unconstrained case without second-order information, but it provides, in the significantly more challenging context studied in our paper, an optimal iteration complexity of $\mathcal{O}(\epsilon^{-2})$ rather than the $\mathcal{O}(\epsilon^{-4})$ obtained in the simpler context only assuming an unbiased oracle with bounded variance.

Assumption AS.5 is weaker than assuming

$$\mathbb{E}_k\left[\|(G_k - g_k)\|^2\right] \leq \kappa_{Gg}^4 \mathbb{E}_k\left[\|s_k\|^2\right],$$

as under this condition we have

$$\mathbb{E}_k\left[|(G_k - g_k)^T s_k|\right] \leq \mathbb{E}_k[\|G_k - g_k\|\|s_k\|] \leq \sqrt{\mathbb{E}_k[\|G_k - g_k\|^2]}\sqrt{\mathbb{E}_k[\|s_k\|^2]} \leq \kappa_{Gg}^2 \mathbb{E}_k\left[\|s_k\|^2\right].$$

In the paper [10] it is shown as in the deterministic case the condition

$$\|(G_k - g_k)\|^2 \leq \kappa_{Gg}^4 \|s_k\|^2,$$

may be ensured by finite differences approximation of the gradient.

The following "linear descent" lemma is a variant of [24, Lemma 2.1] (which, in the unconstrained deterministic context, uses a different optimality measure and a different definition of s_k^L). It is also possible to replace AS.5 with the condition

$$\mathbb{E}_k\left[|(G_k - g_k)^T s_k|\right] \leq \kappa_{Gg}^2 \mathbb{E}_k\left[\|s_k\|^2\right] + Q_k,$$

the rate of convergence then depends on the rate of convergence of $\{Q_k\}$ to zero (see the Appendix).

Lemma 3.1 Suppose that AS.3 and AS.5 hold. Then, for $j \geq 0$,

$$\mathbb{E}_j\left[G_j^T s_j\right] \leq -\frac{\tau\varsigma^2}{2\kappa_B}\,\mathbb{E}_j\left[d_j^T \Delta_j\right] + \kappa_s^2\left(\frac{1}{2}\kappa_B + \kappa_{Gg}^2\right)\mathbb{E}_j\left[\|\Delta_j\|^2\right], \tag{13}$$

where $\Delta_j \in \mathbb{R}^n$ is the vector whose i-th component is $\Delta_{j,i}$.

Proof. Consider any component $i \in \{1, \ldots, n\}$. We first note that (5) and the contractive nature of the projection $P_{\mathcal{F}}$ ensure that $|g_{j,i}| \geq |d_{j,i}|$. Moreover, (5), (6), (7) and (8) implies that either $|s^L_{j,i}| = |d_{j,i}|/w_{j,i}$ or $|s^L_{j,i}| = |d_{j,i}|$. In the first case, we may deduce that

$$|g_{j,i}s^L_{j,i}| \geq \frac{d^2_{j,i}}{w_{j,i}} = w_{j,i}\frac{d^2_{j,i}}{w^2_{j,i}} \geq \varsigma(s^L_{j,i})^2, \tag{14}$$

while, in the latter case,

$$|g_{j,i}s^L_{j,i}| \geq d^2_{j,i} = (s^L_{j,i})^2 = w_{j,i}\frac{d^2_{j,i}}{w_{j,i}} \geq \varsigma\frac{d^2_{j,i}}{w_{j,i}}. \tag{15}$$

Combining the two cases and remembering that $\varsigma \in (0,1]$, we see that, for $i \in \{1, \ldots, n\}$,

$$|g_{j,i}s^L_{j,i}| \geq \varsigma\frac{d^2_{j,i}}{w_{j,i}} \quad \text{and} \quad |g_{j,i}s^L_{j,i}| \geq \varsigma(s^L_{j,i})^2. \tag{16}$$

Summing over all components $i \in \{1, \ldots, n\}$ and using the fact that, by construction, $g_{j,i}s^L_{j,i} < 0$ for all i then gives that

$$g_j^T s_j^L \leq -\varsigma\sum_{i=1}^n \frac{d^2_{j,i}}{w_{j,i}} \quad \text{and} \quad |g_j^T s_j^L| \geq \varsigma\|s_j^L\|^2. \tag{17}$$

We now consider the quadratic model and suppose first that $\gamma_j < 1$. Thus $(s_j^L)^T B_j s_j^L > 0$. We then deduce from (9), AS.3 and (17) that

$$g_j^T s_j^Q + \tfrac{1}{2}(s_j^Q)^T B_j s_j^Q = -\frac{(g_j^T s_j^L)^2}{2(s_j^L)^T B_j s_j^L} \leq -\frac{\varsigma}{2\kappa_B}|g_j^T s_j^L| \leq -\frac{\varsigma^2}{2\kappa_B}\sum_{i=1}^n \frac{d^2_{j,i}}{w_{j,i}}. \tag{18}$$

If now $\gamma_j = 1$, then (9) and (17) give that

$$g_j^T s_j^Q + \frac{1}{2}(s_j^Q)^T B_j s_j^Q = g_j^T s_j^L + \frac{1}{2}(s_j^L)^T B_j s_j^L \leq \frac{1}{2}g_j^T s_j^L < -\frac{\varsigma}{2}\sum_{i=1}^n \frac{d^2_{j,i}}{w_{j,i}}$$

and (18) then again follows from the bounds $\kappa_B \geq 1$ and $\varsigma \leq 1$. Thus, successively using the third part of (10), AS.3, the second part of (10) and (18),

$$\begin{aligned}
g_j^T s_j &= g_j^T s_j + \tfrac{1}{2}s_j^T B_j s_j - \frac{1}{2}s_j^T B_j s_j \\
&\leq \tau\left(g_j^T s_j^Q + \frac{1}{2}(s_j^Q)^T B_j s_j^Q\right) + \frac{1}{2}\kappa_B\|s_j\|^2 \\
&\leq \tau\left(g_j^T s_j^Q + \frac{1}{2}(s_j^Q)^T B_j s_j^Q\right) + \frac{1}{2}\kappa_s^2\kappa_B\|\Delta_j\|^2 \\
&\leq -\frac{\tau\varsigma^2}{2\kappa_B}\sum_{i=1}^n \frac{d^2_{j,i}}{w_{j,i}} + \frac{1}{2}\kappa_s^2\kappa_B\|\Delta_j\|^2,
\end{aligned}$$

which, with the second part of (6), gives that

$$g_j^T s_j \leq -\frac{\tau\varsigma^2}{2\kappa_B}d_j^T\Delta_j + \frac{1}{2}\kappa_s^2\kappa_B\|\Delta_j\|^2. \tag{19}$$

But, using AS.5 and the second part of (10),

$$\begin{aligned}
\mathbb{E}_j\left[G_j^T s_j\right] &= \mathbb{E}_j\left[g_j^T s_j\right] + \mathbb{E}_j\left[(G_j - g_j)^T s_j\right] \\
&\leq \mathbb{E}_j\left[g_j^T s_j\right] + |\mathbb{E}_j\left[(G_j - g_j)^T s_j\right]| \\
&\leq \mathbb{E}_j\left[g_j^T s_j\right] + \kappa_{Gg}^2|\mathbb{E}_j\left[\|s_j\|^2\right]| \\
&\leq \mathbb{E}_j\left[g_j^T s_j\right] + \kappa_{Gg}^2\kappa_s^2\mathbb{E}_j\left[\|\Delta_j\|^2\right]
\end{aligned}$$

and (13) follows from (19). □

Lemma 3.1 is crucial for the proof of our complexity bounds, which also depends on two technical results.

Lemma 3.2 Suppose that, for $t > 0$ and $a, c > 0$,

$$at \leq b + c\,\log(t). \tag{20}$$

Then,

$$t \leq \frac{2b}{a} + \frac{2c}{a}\left[\log\left(\frac{2c}{a}\right) - 1\right]. \tag{21}$$

Proof. The concavity of the logarithm at 2 gives that, for all $x > 0$

$$\log(x) \leq \log(2) + \frac{x-2}{2} = \frac{x}{2} + \log(2) - 1.$$

Applying this inequality for $x = at/c$ then gives that

$$\log(t) = \frac{at}{2c} + \log\left(\frac{2c}{a}\right) - 1,$$

which we may substitute in the inequality $2at \leq 2b + 2c\log(t)$ obtained by multiplying (20) by 2, yielding that

$$2t \leq \frac{2b}{a} + \frac{2c}{a}\left[\frac{at}{2c} + \log\left(\frac{2c}{a}\right) - 1\right] = \frac{2b}{a} + t + \frac{2c}{a}\left[\log\left(\frac{2c}{a}\right) - 1\right],$$

and (21) follows by substracting t on both sides. □

Lemma 3.3 Let $\{a_j\}_{j\geq 0}$ be a sequence of non-negative numbers and let $b_j = \sum_{i=0}^{j} a_i$. Then

$$\sum_{j=0}^{k} \frac{a_j}{\varsigma + b_j} \leq \log\left(1 + \frac{1}{\varsigma} b_k\right).$$

Proof. See [42]. □

Our first complexity result now consider the global rate of convergence of $\|d_k\|$ to zero.

Theorem 3.4 Suppose that AS.1–AS.5 hold and that the ADAGB2 algorithm is applied to problem (1). Then

$$\mathbb{E}\left[\operatorname*{average}_{j\in\{0,\ldots,k\}} \|d_j\|\right] \leq \frac{\kappa_{\rm conv}}{\sqrt{k+1}}, \tag{22}$$

where

$$\kappa_{\rm conv} = \frac{4\kappa_B}{\tau\varsigma^2}\left\{1 + \Gamma_0 + 2n\kappa_*\left[\log\left(\frac{8n\kappa_*\kappa_B}{\tau\varsigma^{5/2}}\right) - 1\right]\right\} \quad \text{with} \quad \kappa_* = \kappa_s^2(\kappa_{Gg}^2 + \tfrac{1}{2}(\kappa_B + L)). \tag{23}$$

Proof. Consider an iteration j and first note that

$$\|s_j\|^2 \le \kappa_s^2 \|\Delta_j\|^2 = \kappa_s^2 \sum_{i=1}^n \frac{d_{j,i}^2}{w_{j,i}^2}$$

because of the second part of (10) and (6). Lemma 3.1, AS.1 and AS.2 then give that

$$\begin{aligned} \mathbb{E}_j[f(x_{j+1})] &\le f(x_j) + \mathbb{E}_j\left[G_j^T s_j\right] + \frac{L}{2}\mathbb{E}_j\left[\|s_j\|^2\right] \\ &\le f(x_j) - \frac{\tau\varsigma^2}{2\kappa_B}\mathbb{E}_j\left[\sum_{i=1}^n \frac{d_{j,i}^2}{w_{j,i}}\right] + \kappa_s^2(\kappa_{Gg}^2 + \tfrac{1}{2}\kappa_B + \tfrac{1}{2}L)\,\mathbb{E}_j\left[\sum_{i=1}^n \frac{d_{j,i}^2}{w_{j,i}^2}\right]. \end{aligned} \tag{24}$$

Defining $\kappa_* = \kappa_s^2(\kappa_{Gg}^2 + \frac{1}{2}\kappa_B + \frac{1}{2}L)$ and using the law of total expectation gives that

$$\mathbb{E}[f(x_{j+1})] \le \mathbb{E}[f(x_j)] - \frac{\tau\varsigma^2}{2\kappa_B}\mathbb{E}\left[\sum_{i=1}^n \frac{d_{j,i}^2}{w_{j,i}}\right] + \kappa_*\mathbb{E}\left[\sum_{i=1}^n \frac{d_{j,i}^2}{w_{j,i}^2}\right],$$

and therefore, summing for $j \in \{0,\ldots,k\}$ for k fixed, that

$$\frac{\tau\varsigma^2}{2\kappa_B}\mathbb{E}\left[\sum_{j=0}^k\sum_{i=1}^n \frac{d_{j,i}^2}{w_{j,i}}\right] \le f(x_0) - f_{\rm low} + \kappa_*\mathbb{E}\left[\sum_{j=0}^k\sum_{i=1}^n \frac{d_{j,i}^2}{w_{j,i}^2}\right]. \tag{25}$$

Observe now that, if $p = \mathrm{argmax}_{q\in\{1,\ldots,n\}}\, w_{j,q}$,

$$w_{j,i} \le \max_{q\in\{1,\ldots,n\}} w_{j,q} = w_{j,p} = \sqrt{\varsigma + \sum_{\ell=0}^j d_{\ell,p}^2} \le \sqrt{\varsigma + \sum_{\ell=0}^j \|d_\ell\|^2}$$

for all $i \in \{1,\ldots,n\}$ and thus that

$$\sum_{j=0}^k \frac{\|d_j\|^2}{\sqrt{\varsigma + \sum_{\ell=0}^j \|d_\ell\|^2}} \le \sum_{j=0}^k\sum_{i=1}^n \frac{d_{j,i}^2}{\max_{q\in\{1,\ldots,n\}} w_{j,q}} \le \sum_{j=0}^k\sum_{i=1}^n \frac{d_{j,i}^2}{w_{j,i}}. \tag{26}$$

We may now apply Lemma 3.3 to derive that, for each $i \in \{1,\ldots,n\}$,

$$\begin{aligned} \mathbb{E}\left[\sum_{i=1}^n\sum_{j=0}^k \frac{d_{j,i}^2}{w_{j,i}^2}\right] &= \mathbb{E}\left[\sum_{i=1}^n\sum_{j=0}^k \frac{d_{j,i}^2}{\varsigma + \sum_{\ell=0}^j d_{j,\ell}^2}\right] \\ &\le n\mathbb{E}\left[\max_{i\in\{1,\ldots,n\}} \log\left(1 + \frac{1}{\varsigma}\sum_{j=0}^k d_{j,i}^2\right)\right] \\ &\le n\mathbb{E}\left[\log\left(1 + \frac{1}{\varsigma}\sum_{i=1}^n\sum_{j=0}^k d_{j,i}^2\right)\right], \end{aligned}$$

and hence, using AS.4 to define $\Gamma_0 = f(x_0) - f_{\rm low} > 0$, (25) and (26) gives that

$$\frac{\tau\varsigma^2}{2\kappa_B}\mathbb{E}\left[\sum_{j=0}^k \frac{\|d_j\|^2}{\sqrt{\varsigma + \sum_{\ell=0}^j \|d_\ell\|^2}}\right] \le \Gamma_0 + n\kappa_*\mathbb{E}\left[\log\left(1 + \frac{1}{\varsigma}\sum_{j=0}^k \|d_j\|^2\right)\right] \tag{27}$$

in turn ensuring that

$$
\begin{aligned}
\frac{\tau\varsigma^2}{2\kappa_B}\mathbb{E}\left[\sum_{j=0}^k \frac{\|d_j\|^2}{\sqrt{\varsigma+\sum_{\ell=0}^k\|d_\ell\|^2}}\right] &\le \Gamma_0 + n\kappa_*\mathbb{E}\left[\log\left(1+\frac{1}{\varsigma}\sum_{j=0}^k\|d_j\|^2\right)\right] \\
&= \Gamma_0 + 2n\kappa_*\mathbb{E}\left[\log\left(\sqrt{1+\frac{1}{\varsigma}\sum_{j=0}^k\|d_j\|^2}\right)\right] \qquad (28)\\
&\le \Gamma_0 + 2n\kappa_*\log\left(\mathbb{E}\left[\sqrt{1+\frac{1}{\varsigma}\sum_{j=0}^k\|d_j\|^2}\right]\right),
\end{aligned}
$$

where we used Jensen's inequality and the concavity of the logarithm to deduce the last inequality. Now

$$
\sum_{j=0}^k \frac{\|d_j\|^2}{\sqrt{\varsigma+\sum_{\ell=0}^k\|d_\ell\|^2}} = \sqrt{\varsigma+\sum_{\ell=0}^k\|d_\ell\|^2} - \frac{\varsigma}{\sqrt{\varsigma+\sum_{\ell=0}^k\|d_\ell\|^2}}
$$

and thus

$$
\sqrt{\varsigma}\,\mathbb{E}\left[\sqrt{1+\frac{1}{\varsigma}\sum_{\ell=0}^k\|d_\ell\|^2}\right] = \mathbb{E}\left[\sqrt{\varsigma+\sum_{\ell=0}^k\|d_\ell\|^2}\right] \le \mathbb{E}\left[\sum_{j=0}^k\frac{\|d_j\|^2}{\sqrt{\varsigma+\sum_{\ell=0}^k\|d_\ell\|^2}}\right]+\sqrt{\varsigma}.
$$

Substituting (28) in this inequality and using that $\tau\varsigma^2\sqrt{\varsigma}\le 1 < 2\kappa_B$ then gives that

$$
\frac{\tau\varsigma^2\sqrt{\varsigma}}{2\kappa_B}\mathbb{E}\left[\sqrt{1+\frac{1}{\varsigma}\sum_{\ell=0}^k\|d_\ell\|^2}\right] \le 1+\Gamma_0+2n\kappa_*\log\left(\mathbb{E}\left[\sqrt{1+\frac{1}{\varsigma}\sum_{\ell=0}^k\|d_\ell\|^2}\right]\right),
$$

which can be rewritten as

$$
a\,t_k \le b + c\,\log(t_k)
$$

with

$$
t_k = \mathbb{E}\left[\sqrt{1+\frac{1}{\varsigma}\sum_{\ell=0}^k\|d_\ell\|^2}\right]\ge 1, \quad a=\frac{\tau\varsigma^{\frac{5}{2}}}{2\kappa_B}, \quad b = 1+\Gamma_0 \quad \text{and} \quad c = 2n\kappa_*.
$$

We may then use Lemma 3.2 and deduce that

$$
t_k \le \frac{2b}{a}+\frac{2c}{a}\left[\log\left(\frac{2c}{a}\right)-1\right] = \frac{4\kappa_B}{\tau\varsigma^{5/2}}\left\{1+\Gamma_0+2n\kappa_*\left[\log\left(\frac{8n\kappa_*\kappa_B}{\tau\varsigma^{5/2}}\right)-1\right]\right\},
$$

which gives that

$$
\begin{aligned}
\mathbb{E}\left[\sqrt{\sum_{\ell=0}^k\|d_\ell\|^2}\right] &\le \sqrt{\varsigma}\,\mathbb{E}\left[\sqrt{1+\frac{1}{\varsigma}\sum_{\ell=0}^k\|d_\ell\|^2}\right] \\
&\le \frac{4\kappa_B}{\tau\varsigma^2}\left\{1+\Gamma_0+2n\kappa_*\left[\log\left(\frac{8n\kappa_*\kappa_B}{\tau\varsigma^{5/2}}\right)-1\right]\right\} \\
&= \kappa_{\rm conv}.
\end{aligned}
$$

Dividing by $\sqrt{k+1}$ and using the inequality

$$
\frac{1}{k+1}\sum_{j=0}^k\|d_j\| \le \frac{1}{\sqrt{k+1}}\sqrt{\sum_{j=0}^k\|d_j\|^2}
$$

finally gives

$$\mathbb{E}\left[\operatorname*{average}_{j\in\{0,\ldots,k\}}\|d_j\|\right] \leq \mathbb{E}\left[\sqrt{\operatorname*{average}_{j\in\{0,\ldots,k\}}\|d_j\|^2}\right] \leq \frac{\kappa_{\rm conv}}{\sqrt{k+1}},$$

as requested. □

In the deterministic case, Theorem 3.4 provides a bound on the complexity of solving the bound-constrained problem (1) for which we consider the standard optimality measure [13, Section 12.1] $\|\Xi_k\|$, where

$$\Xi_k \stackrel{\rm def}{=} P_{\mathcal{F}}(x_k - G_k) - x_k$$

(if the problem is unconstrained, $\|\Xi_k\| = \|G_k\|$).

Corollary 3.5 Suppose that AS.1–AS.4 hold, that the ADAGB2 algorithm is applied to problem (1) and that $g_j = G_j$ for all j. Then

$$\min_{j\in\{0,\ldots,k\}} \|\Xi_j\| \leq \operatorname*{average}_{j\in\{0,\ldots,k\}} \|\Xi_j\| \leq \frac{\kappa_{\rm conv}}{\sqrt{k+1}}, \tag{29}$$

where the constant $\kappa_{\rm conv}$ is computed as in Theorem 3.4 using the value $\kappa_{Gg} = 0$.

Proof. The result trivially follows from Theorem 3.4 with $\kappa_{Gg} = 0$. □

4 First-order optimality and stochastic projected gradients with bounds

4.1 A more general framework

Corollary 3.5 gives the global rate of convergence of the ADAGB2 algorithm in the deterministic case where (obviously) $G_k = g_k$ and thus $\Xi_k = d_k$. Things are more complicated in the stochastic case because, as we prove in this section, this last inequality may no longer hold. Thus, while Theorem 3.4 gives the rate of convergence of the *approximate* criticality measure, obtaining a similar rate for the *true* criticality measure is another question. Because our way to answer this question applies not only to the ADAGB2 algorithm, but also to a wider class of stochastic projected-gradient methods, we now consider solving problem (1) using the algorithmic framework given by StochProjGrad.

Algorithm 4.1: StochProjGrad $(x_{\rm ini}, l, u)$

Step 0: Initialization: Set $x_0 = P_{\mathcal{F}}(x_{\rm ini})$, $k = 0$.

Step 1: Step: Compute $g_k = g(x_k, \xi)$ a random approximation of $G_k = \nabla_x^1 f(x_k)$, and a step s_k such that $x_k + s_k \in \mathcal{F}$,

Step 4: Loop: Set $x_{k+1} = x_k + s_k$, increment k by one and go to Step 1.

Moreover we will assume, in this subsection, that the expected approximate criticality measure

$$\mathbb{E}_k[\|d_k\|] = \mathbb{E}_k[\|P_{\mathcal{F}}(x_k - g_k) - x_k\|]$$

converges to zero (at least on average as in Theorem 3.4). We are now interested in what can be deduced on $\mathbb{E}[\|\Xi_k\|] = \mathbb{E}[\|P_{\mathcal{F}}(x_k - G_k) - x_k\|]$, the relevant criticality measure for problem (1) in the stochastic case. Ideally, one would hope that the approximate gradient's distribution ensures coherence of the measures in the sense that

$$\mathbb{E}[\|\Xi_k\|] \leq \kappa_{\text{opt}} \mathbb{E}[\|d_k\|] \tag{30}$$

for a fixed $\kappa_{\text{opt}} > 0$, in which case $\mathbb{E}[\|\Xi_k\|]$ converges to zero at the same rate as $\mathbb{E}[\|d_k\|]$. If we consider the unconstrained case ($\mathcal{F} = \mathbb{R}^n$) and assume that the gradient oracle is unbiased (i.e. $\mathbb{E}_k[g_k] = G_k$), then, using Jensen's inequality and the convexity of the norm,

$$\|\Xi_k\| = \|G_k\| = \|\mathbb{E}_k[g_k]\| \leq \mathbb{E}_k[\|g_k\|] = \mathbb{E}_k[\|d_k\|].$$

Taking the full expectation on both sides and using the law of total expectation then shows that (30) always holds with $\kappa_{\text{opt}} = 1$. The situation is very different if bounds are present, even if the gradient oracle is unbiased. To see this, consider the following one dimensional example, where $\mathcal{F} = [0, +\infty)$ and

$$x_k = \frac{1}{k+1} \quad \text{and} \quad g_k = \begin{cases} 1 & \text{with probability } \; p_k = \dfrac{1}{k+1} + \dfrac{1}{(k+1)^2} \\ 0 & \text{with probability } \; 1 - p_k, \end{cases}$$

for $k > 0$. Also define $G_k = \mathbb{E}_k[g_k] = p_k$ (so that the gradient oracle is unbiased). One then easily verifies that $|G_{k+1} - G_k| \leq 3\,|x_{k+1} - x_k|$, so that AS.2 holds with $L = 3$. Indeed, since $G_k = x_k + x_k^2$, we have that

$$|G_{k+1} - G_k| \leq |x_{k+1} - x_k + x_{k+1}^2 - x_k^2| \leq |x_{k+1} - x_k + (x_{k+1} - x_k)(x_{k+1} + x_k)| \leq 3\,|x_{k+1} - x_k|.$$

These definitions give that

$$|\Xi_k| = \left| P_{[0,+\infty)}\left(\frac{1}{k+1} - p_k\right) - \frac{1}{k+1}\right| = \left| P_{[0,+\infty)}\left(\frac{-1}{(k+1)^2}\right) - \frac{1}{k+1}\right| = \frac{1}{k+1}$$

and

$$\begin{aligned} \mathbb{E}_k[|d_k|] &= \left| P_{[0,+\infty)}\left(\tfrac{1}{k+1} - 1\right) - \tfrac{1}{k+1}\right| p_k + \left| P_{[0,+\infty)}\left(\tfrac{1}{k+1} - 0\right) - \tfrac{1}{k+1}\right| (1 - p_k) \\ &= \frac{1}{k+1}\, p_k + 0\,(1 - p_k) \\ &= \frac{1}{(k+1)^2} + \frac{1}{(k+1)^3}. \end{aligned}$$

Applying now the law of total expectation, we see that

$$\lim_{k\to\infty} \frac{\mathbb{E}[|d_k|]}{\mathbb{E}[|\Xi_k|]} = 0,$$

preventing (30) (a "coherently distributed" gradient oracle) to hold for a fixed $\kappa_{\text{opt}} > 0$. We therefore conclude that, in general, *the rate of decrease of* $\mathbb{E}[\|d_k\|]$ *does* not *translate to a similar rate of decrease for* $\mathbb{E}[\|\Xi_k\|]$*, even for unbiased gradient oracles.*

Fortunately, the situation can be improved by strengthening the condition on the gradient accuracy, even if the gradient oracle is biased. This is the object of the next lemma.

Lemma 4.1 For each $k \geq 0$ and each $i \in \{1, \ldots, n\}$, we have that

$$\mathbb{E}[\|\Xi_k\|] \leq \mathbb{E}[\|d_k\|] + \mathbb{E}[\|g_k - G_k\|]\,. \tag{31}$$

Moreover, if $\mathbb{E}_k[\|g_k - G_k\|] \leq \kappa_{\rm err}\,\mathbb{E}_k[\|d_k\|]$ for some $\kappa_{\rm err} \geq 0$, then

$$\mathbb{E}[\|\Xi_k\|] \leq (1 + \kappa_{\rm err})\,\mathbb{E}[\|d_k\|]\,. \tag{32}$$

Proof. Consider an arbitrary $i \in \{1, \ldots, n\}$ and note that

$$\Xi_{k,i} = P_i(x_{k,i} - G_{k,i}) - x_{k,i}. \tag{33}$$

where $P_i = P_{[l_i, u_{\ell_i}]}$. Since P_i is a contractive map, we have that

$$\begin{aligned}
&(P_i(x_{k,i} - G_{k,i}) - x_{k,i})^2 - (P_i(x_{k,i} - g_{k,i}) - x_{k,i})^2 \\
&\qquad = [P_i(x_{k,i} - G_{k,i}) - x_{k,i} + P_i(x_{k,i} - g_{k,i}) - x_{k,i}] \times \\
&\qquad\qquad [P_i(x_{k,i} - G_{k,i}) - P_i(x_{k,i} - g_{k,i})] \\
&\qquad \leq |P_i(x_{k,i} - G_{k,i}) - x_{k,i} + P_i(x_{k,i} - g_{k,i}) - x_{k,i}|\; |g_{k,i} - G_{k,i}|.
\end{aligned}$$

Now, again using the contractivity of P_i,

$$\begin{aligned}
&|P_i(x_{k,i} - G_{k,i}) - x_{k,i} + P_i(x_{k,i} - g_{k,i}) - x_{k,i}| \\
&\qquad = \big|2(P_i(x_{k,i} - g_{k,i}) - x_{k,i}) \\
&\qquad\qquad + (P_i(x_{k,i} - G_{k,i}) - x_{k,i}) - (P_i(x_{k,i} - g_{k,i}) - x_{k,i})\big| \\
&\qquad \leq 2\,|P_i(x_{k,i} - g_{k,i}) - x_{k,i}| + |P_i(x_{k,i} - G_{k,i}) - P_i(x_{k,i} - g_{k,i})| \\
&\qquad \leq 2\,|P_i(x_{k,i} - g_{k,i}) - x_{k,i}| + |g_{k,i} - G_{k,i}|
\end{aligned}$$

and thus

$$\begin{aligned}
&(P_i(x_{k,i} - G_{k,i}) - x_{k,i})^2 - (P_i(x_{k,i} - g_{k,i}) - x_{k,i})^2 \\
&\qquad \leq 2\,|P_i(x_{k,i} - g_{k,i}) - x_{k,i}|\; |g_{k,i} - G_{k,i}| + |g_{k,i} - G_{k,i}|^2\,,
\end{aligned}$$

which, using (5) and (33), gives that

$$\Xi_{k,i}^2 \leq d_{k,i}^2 + 2|d_{k,i}|\,|g_{k,i} - G_{k,i}| + |g_{k,i} - G_{k,i}|^2$$

and therefore, summing for $i \in \{1, \ldots, n\}$ and using the Cauchy-Schwartz inequality, that

$$\begin{aligned}
\|\Xi_k\|^2 &= \|d_k\|^2 + 2|d_k|^T|g_k - G_k| + \|g_k - G_k\|^2 \\
&\leq \|d_k\|^2 + 2\|d_k\|\,\|g_k - G_k\| + \|g_k - G_k\|^2 \\
&= \Big(\|d_k\| + \|g_k - G_k\|\Big)^2.
\end{aligned}$$

This yields that

$$\|\Xi_k\| \leq \|d_k\| + \|g_k - G_k\|. \tag{34}$$

and (31) follows by taking conditional expectations on both sides. Moreover, if $\mathbb{E}_k[\|g_k - G_k\|] \leq \kappa_{\rm err}\,\mathbb{E}_k[\|d_k\|]$, (32) results from (34) and the tower property. □

Thus we see from (31) that the convergence of $\mathbb{E}[\|\Xi_k\|]$ to zero is still guaranteed if $\mathbb{E}[\|g_k - G_k\|]$ converges to zero along with $\mathbb{E}[\|d_k\|]$. Moreover, if it does so at the same speed, the rate of decrease of $\mathbb{E}[\|d_k\|]$ dominates, as shown by (32).

This result also indicates what can happen if nothing is known on the error in the gradient. The optimality measure of the approximate problem goes to zero (as in Theorem 3.4) but the true measure could remain at a level given by the second term of (31). To interpret this term, observe that, using Jensen's inequality and the concavity of the square root,

$$\mathbb{E}_k[\|g_k - G_k\|] = \mathbb{E}_k\left[\sqrt{\|g_k - G_k\|^2}\right] \le \sqrt{\mathbb{E}_k[\|g_k - G_k\|^2]} = \text{RMSE}_k \tag{35}$$

for all $k \ge 0$, where RMSE_k is the conditional root mean square error (RMSE) of the gradient oracle at iteration k. Thus, by the law of total expectation, $\mathbb{E}[\|g_k - G_k\|] \le \mathbb{E}[\text{RMSE}_k]$, and $\limsup_{k\to\infty} \mathbb{E}[\text{RMSE}_k]$ gives an asymptotic upper bound on the criticality default (the deviation of the criticality measure from zero).

4.2 Application to the ADAGB2 algorithm

The discussion of the previous subsection finally allows us to rephrase Theorem 3.4 to cover convergence of the ADAGB2 algorithm on problem (1) in three progressively more general scenarii (condition (37) below was called "new2" in Table 1).

Theorem 4.2 Suppose that AS.1–AS.5 hold and that the ADAGB2 algorithm is applied to problem (1). Then

$$\underset{j\in\{0,\dots,k\}}{\text{average}}\ \mathbb{E}[\|\Xi_j\|] \le \frac{\kappa_{\text{conv}}\kappa_1}{\sqrt{k+1}} + \kappa_2 \underset{j\in\{0,\dots,k\}}{\text{average}}\ \mathbb{E}[\|g_k - G_k\|] + \kappa_3 \tag{36}$$

where κ_{conv} is defined in (23) and

coherently distributed: $\kappa_1 = \kappa_{\text{opt}}$ and $\kappa_2 = \kappa_3 = 0$ if

$$\mathbb{E}[\|\Xi_k\|] \le \kappa_{\text{opt}}\mathbb{E}[\|d_k\|]$$

for some constant $\kappa_{\text{opt}} > 0$ and all $k \ge 0$;

controlled error: $\kappa_1 = 1 + \kappa_{\text{err1}}$ and $\kappa_2 = \kappa_3 = 0$ if

$$\mathbb{E}_k[\|g_k - G_k\|] \le \kappa_{\text{err1}}\,\mathbb{E}_k[\|d_k\|] \tag{37}$$

for some constant $\kappa_{\text{err1}} > 0$ and all $k \ge 0$;

probabilistic error control: $\kappa_1 = (1 + \kappa_{\text{pec}})$, $\kappa_2 = 0$ and $\kappa_3 = (1-p)\kappa_{\text{err2}}$ if

$$\big\|g_k - G_k\big\| < \kappa_{\text{err2}} \tag{38}$$

$$\mathbb{P}_k\big[\mathcal{E}_k\big] \ge p \text{ with event } \mathcal{E}_k = \{\big\|g_k - G_k\big\| \le \kappa_{\text{pec}}\|d_k\|\} \tag{39}$$

for some $p \in (0,1)$, $\kappa_{\text{err2}}, \kappa_{\text{pec}} \ge 0$ and all $k \ge 0$;

general: $\kappa_1 = \kappa_2 = 1$, $\kappa_3 = 0$, otherwise.

Proof. The results for the "coherently distributed", "controlled error" and "general" scenarii are obtained by combining Theorem 3.4 with Lemma 4.1 and (for the "coherently distributed" and "controlled error" scenarii) from the comments following it.

For analyzing the "probabilistic error control" scenario, first note that, using the contractive nature of projection,

$$\|\Xi_k\| = \|P_{\mathcal{F}}(x_k - g_k) - x_k + P_{\mathcal{F}}(x_k - G_k) - P_{\mathcal{F}}(x_k - g_k)\| \leq \|d_k\| + \|g_k - G_k\|.$$

If $\mathbb{1}(\mathcal{E}_k) = 1$ then this inequality implies that

$$\|\Xi_k\| \leq (\kappa_{\rm pec} + 1)\|d_k\|,$$

and it also implies that in all cases,

$$\|\Xi_k\| \leq \|d_k\| + \kappa_{\rm err2} \leq (\kappa_{\rm pec} + 1)\|d_k\| + \kappa_{\rm err2}.$$

Thus, letting $\overline{\mathcal{E}_k} = \{\|g_k - G_k\| > \kappa_{\rm pec}\|d_k\|\}$ be the complement of $\mathcal{E}$, we deduce that

$$\begin{aligned}\mathbb{E}_k[\|\Xi_k\|] &= P_k(\mathcal{E}_k)\,\mathbb{E}_k[\|\Xi_k\| \,|\, \mathcal{E}_k] + P_k(\overline{\mathcal{E}_k})\,\mathbb{E}_k[\|\Xi_k\| \,|\, \overline{\mathcal{E}_k}]\\ &\leq (\kappa_{\rm pec} + 1)\mathbb{E}_k[\|d_k\|] + (1-p)\kappa_{\rm err2}.\end{aligned}$$

Using now the tower property and (22) we obtain that

$$\operatorname*{average}_{j\in\{0,\ldots,k\}} \mathbb{E}\|\Xi_j\| \leq \frac{\kappa_{\rm conv}(\kappa_{\rm pec}+1)}{\sqrt{k+1}} + (1-p)\kappa_{\rm err2}.$$

□

If the function f is subject to independent and identically distributed noise, then gradient can be estimated by standard sample averaging approximation techniques and the probabilistic condition in (39) can be enforced by the Bernstein inequality [4, 34]. The term depending on inaccurate approximations of the gradient is damped by the factor $(1-p)$ and its magnitude can be made arbitrarily small by increasing p.

Theorem 4.2 gives the desired fast convergence to zero of the first-order optimality measure associated with problem (1) in the coherently distributed and controlled error cases. In particular, if the problem is unconstrained and the gradient oracle is unbiased, then the coherently distributed case applies with $\kappa_{\rm opt} = 1$ and we *recover the* $\mathcal{O}(\epsilon^{-2})$ *complexity bound for* AdaGrad obtained in the references mentioned in the introduction *assuming the condition given by AS.5.* The same is true in the bound-constrained case in the controlled-error scenario.

Theorem 4.2 and the discussion of the previous section also indicate that the true measure could remain at a level given by $\limsup_{k\to\infty} \beta_k$ where $\beta_k = \text{average}_{j\in\{0,\ldots,k\}}\, \mathbb{E}[\|g_k - G_k\|]$ is bounded above by the average RMSE_k of the gradient oracle, the average being taken on the first k iterations. If β_k is bounded for large k by a (hopefully) small positive constant, then the optimality measure is only constrained to fall below this constant. However, convergence to true optimality must still happen if the sequence $\{\beta_k\}$ converges to zero[3], albeit possibly at a rate slower than $\mathcal{O}(1/\sqrt{k+1})$.

Moreover, reformulating the "new2" condition (37) using the bound (35) allows us to state the following direct consequence of Theorem 4.2.

Corollary 4.3 Suppose that AS.1–AS.5 hold and that the ADAGB2 algorithm is applied to problem (1). Suppose also that there exists a constant $\kappa_{\rm err} \geq 0$ such that, for all $k \geq 0$,

$$\text{RMSE}_k \leq \kappa_{\rm err}\, \mathbb{E}_k[\|d_k\|]$$

where RMSE_k is the conditional root mean square error defined by the last equality of (35). Then

$$\operatorname*{average}_{j\in\{0,\ldots,k\}} \mathbb{E}[\|\Xi_j\|] \leq \frac{(1+\kappa_{\rm err})\kappa_{\rm conv}}{\sqrt{k+1}},$$

where $\kappa_{\rm conv}$ defined by (23).

[3] Which does not necessarily requires the convergence of the sequence $\{\|g_k - G_k\|\}$ to zero.

We finally state a complexity result in probability simply derived from Theorem 4.2 for its first two scenarii.

Corollary 4.4 Suppose that the conditions of Theorem 4.2 hold for the coherently distributed or controlled-error cases. Then, given $\delta \in (0,1)$, one has that

$$\mathbb{P}\left[\min_{j\in\{0,\ldots,k\}} \|\Xi_j\| \leq \epsilon\right] \geq 1-\delta \quad \text{for} \quad k > \left(\frac{\kappa_1 \kappa_{\rm conv}}{\delta\epsilon}\right)^2,$$

where $\kappa_{\rm conv}$ is given by (23).

Proof. Using Markov's inequality and (36) with $\kappa_2 = \kappa_3 = 0$, we have that

$$\mathbb{P}\left[\min_{j\in\{0,\ldots,k\}} \|\Xi_j\| \leq \epsilon\right] \geq \mathbb{P}\left[\frac{1}{k}\sum_{j=0}^{k} \|\Xi_j\| \leq \epsilon\right] \geq 1 - \frac{1}{\epsilon}\,\mathbb{E}\left[\frac{1}{k}\sum_{j=0}^{k}\|\Xi_j\|\right] \geq 1 - \frac{\kappa_1\kappa_{\rm conv}}{\epsilon\sqrt{k+1}}.$$

The desired conclusion follows. □

5 Conclusions and perspectives

We have introduced the ADAGB2 algorithm for bound-constrained stochastic optimization which is also capable of using available second-order information. When second-order information is not used and the problem is unconstrained, ADAGB2 subsumes the AdaGrad algorithm.

We have shown that, given $\delta \in (0,1)$ and $\epsilon \in (0,1]$, *the* ADAGB2 *algorithm needs at most* $\mathcal{O}(\epsilon^{-2})$ *iterations to ensure an* ϵ*-approximate first-order critical point of the bound-constrained problem (1) with probability at least* $1-\delta$, provided that Assumption AS.5 holds and the average RMSE β_k is sufficiently small. Should this condition fail, we have shown that the optimality default is bounded above for large k by the limit superior of the average oracle's RMSE.

We have also discussed the relation between the standard optimality measure for bound-constrained induced by the projected gradient in the stochastic case, and have shown that, in general, the unbiased nature of the gradient oracle is not sufficient to ensure convergence on the true constrained problem at the optimal rate, even if such a convergence occurs for the approximate one.

Some interesting theoretical questions remain open at this point, including the extension of the results to (L_0, L_1)-smooth functions, the use of momentum, ensuring second-order optimality and the handling of more general constraints.

Acknowledgements

Serge Gratton and Philippe Toint are grateful to Defeng Sun, Xiaojun Chen and Zaikun Zhang of the Department of Applied Mathematics of the Hong-Kong Polytechnic University for their support during a research visit in the fall 2024. Philippe Toint also acknowledges the support of DIEF (UNIFI) for a visit in October 2024. The research of Stefania Bellavia and Benedetta Morini was partially supported by INDAM-GNCS through Progetti di Ricerca 2023 and by PNRR - Missione 4 Istruzione e Ricerca - Componente C2 Investimento 1.1, Fondo per il Programma Nazionale di Ricerca e Progetti di Rilevante Interesse Nazionale (PRIN) funded by the European Commission under the NextGeneration EU programme, project "Advanced optimization METhods for automated central veIn Sign detection in multiple sclerosis from magneTic resonAnce imaging (AMETISTA)", code: P2022J9SNP, MUR D.D. financing decree n. 1379 of 1st September 2023 (CUP E53D23017980001), project "Numerical Optimization with Adaptive Accuracy and Applications to Machine Learning", code: 2022N3ZNAX MUR D.D. financing decree n. 973 of 30th June 2023 (CUP B53D23012670006), and by Partenariato esteso FAIR "Future Artificial Intelligence Research" SPOKE 1 Human-Centered AI. Obiettivo 4, Project "Mathematical and Physical approaches to innovative Machine Learning technologies (MaPLe)", Codice Identificativo EP_FAIR_002, CUP B93C23001750006.

The authors are indebted to an anonymous referee for a question that led to a cleaner and more complete statement of the rate of convergence results.

References

[1] A. Alacaoglu and H. Lyu. Convergence of first-order methods for constrained nonconvex optimization with dependent data. In *International Conference on Machine Learning. PMLR*, pages 458–489, 2023.

[2] A. Alacaoglu, Y. Malitsky, and V. Cevher. Convergence of adaptive algorithms for constrained weakly convex optimization. In *Advances in Neural Information Processing Systems, 34*, 2021.

[3] A. Attia and T. Koren. SGD with AdaGrad stepsizes: full adaptivity with high probability to unknown parameters, unbounded gradients and affine variance. In *International Conference on Machine Learning*, 2023.

[4] S. Bellavia and G. Gurioli. Stochastic analysis of an adaptive cubic regularization method under inexact gradient evaluations and dynamic Hessian accuracy. *Optimization*, 71:227–261, 2022.

[5] S. Bellavia, B. Morini, and M. Yousefi. Fully stochastic trust-region methods with Barzilai-Borwein steplengths. *Journal of Computational and Applied Mathematics*, 476:117059, 2026.

[6] A. Berahas, L. Cao, and K. Scheinberg. Global convergence rate analysis of a generic line search algorithm with noise. *SIAM Journal on Optimization*, 31:1489–1518, 2021.

[7] A.S. Berahas, M. Jahani, P. Richtárik, and M. Takáč. Quasi-Newton methods for machine learning: forget the past, just sample. *Optimization Methods and Software*, 37(5):1668–1704, 2022.

[8] J. Blanchet, C. Cartis, M. Menickelly, and K. Scheinberg. Convergence rate analysis of a stochastic trust region method via supermartingales. *INFORMS Journal on Optimization*, 1(2):92–119, 2019.

[9] L. Bottou, F. E. Curtis, and J. Nocedal. Optimization methods for large-scale machine learning. *SIAM Review*, 60(2):223–311, 2018.

[10] Ph. L. Toint C. Cartis, N. I. M. Gould. On the oracle complexity of first-order and derivative-free algorithms for smooth nonconvex minimization. *SIAM Journal on Optimization*, 22(1):66–86, 2012.

[11] S. Cai, Z. Mao, Z. Wang, M. Yin, and G. E. Karniadakis. Physics-informed neural networks (PINNs) for fluid mechanics: A review. *Acta Mechanica Sinica*, 37(12):1727–1738, 2021.

[12] G. Ciaramella, F. Nobile, and T. Vanzan. A multigrid solver for PDE-constrained optimization with uncertain inputs. *Journal of Scientific Computing*, 101(1):13, 2024.

[13] A. R. Conn, N. I. M. Gould, and Ph. L. Toint. *Trust-Region Methods.* Number 1 in MOS-SIAM Optimization Series. SIAM, Philadelphia, USA, 2000.

[14] F. Curtis, V. Kungurtsev, D. Robinson, and Q. Wang. A stochastic-gradient-based interior-point algorithm for solving smooth bound-constrained optimization problems, 2025.

[15] F.E. Curtis and R. Shi. A fully stochastic second-order trust region method. *Optimization Methods and Software*, 37(3):844–877, 2022.

[16] Damek Davis and Dmitriy Drusvyatskiy. Stochastic model-based minimization of weakly convex functions. *SIAM Journal on Optimization*, 29(1):207–239, 2019.

[17] J. de Faria, R. Assunção, and F. Murai. Fisher scoring method for neural networks optimization. In *SIAM International Conference on Data Mining (SDM)*, pages 748–756. SIAM, 2023.

[18] A. Défossez, L. Bottou, F. Bach, and N. Usunier. A simple convergence proof for Adam and Adagrad. *Transactions on Machine Learning Research*, October 2022.

[19] J. Duchi, E. Hazan, and Y. Singer. Adaptive subgradient methods for online learning and stochastic optimization. *Journal of Machine Learning Research*, 12(7):2121–2159, 2011.

[20] M. Faw, L. Rout, C. Caramanis, and S. Shakkottai. Beyond uniform smoothness: a stopped analysis of adaptive SGD. In *Conference on Learning Theory*, 2023.

[21] M. Faw, I. Tziotis, C. Caramanis, A. Mokhtari, S. Shakkottai, and R. Ward. The power of adaptivity in SGD: self-tuning step sizes with unbounded gradients and affine variance. In *Conference on Learning Theory*, 2022.

[22] S. Ghadimi, G. Lan, and H. Zhang. Mini-batch stochastic approximation methods for nonconvex stochastic composite optimization. *Mathematical Programming*, 155(1):267–305, 2016.

[23] S. Gratton, S. Jerad, and Ph. L. Toint. First-order objective-function-free optimization algorithms and their complexity. arXiv:2203.01757, 2022.

[24] S. Gratton, S. Jerad, and Ph. L. Toint. Complexity of Adagrad and other first-order methods for nonconvex optimization problems with bounds constraints. arXiv:2406.15793, 2024.

[25] Y. Hong and J. Lin. Revisting convergence of Adagrad with relaxed assumptions. In *UAI '24: Proceedings of the Fortieth Conference on Uncertainty in Artificial Intelligence*, 2024.

[26] R. Jiang, D. Maladkar, and A. Mokhtari. Convergence analysis of adaptive gradient methods under refined smoothness and noise assumptions. arXiv:2406.04592, 2024.

[27] A. Kavis, K. Y. Levy, and V. Cevher. High probability bounds for a class of nonconvex algorithms with AdaGrad stepsize. In *International Conference on Learning Representations*, 2022.

[28] A. Kopaničáková and R. Krause. A recursive multilevel trust region method with application to fully monolithic phase-field models of brittle fracture. *Computer Methods in Applied Mechanics and Engineering*, 360:112720, 2020.

[29] R. Krause. A Nonsmooth Multiscale Method for Solving Frictional Two-Body Contact Problems in 2D and 3D with Multigrid Efficiency. *SIAM Journal on Scientific Computing*, 31(2):1399–1423, 2009.

[30] B. Leimkuhler, T. J. Vlaar, T. Pouchon, and A. Storkey. Better training using weight-constrained stochastic dynamics. In *International Conference on Machine Learning*, pages 6200–6211. PMLR, 2021.

[31] X. Li and F. Orabona. On the convergence of stochastic gradient descent with adaptive stepsizes. In *The 22nd International Conference on Artificial Intelligence and Statistics*, page 983–992, 2019.

[32] Z. Liu, T. D. Nguyen, A. Ene, and H. Nguyen. On the convergence of Adagrad(Norm) on $\mathbb{R}^d$: Beyond convexity, non-asymptotic rate and acceleration. In *International Conference on Learning Representations*, 2023.

[33] L. Lu, R. Pestourie, W. Yao, Z. Wang, F. Verdugo, and S. G. Johnson. Physics-informed neural networks with hard constraints for inverse design. *SIAM Journal on Scientific Computing*, 43(6):B1105–B1132, 2021.

[34] P. G. Martinsson and J. A. Tropp. Randomized numerical linear algebra: Foundations and algorithms. *Acta Numerica*, 29:403–572, 2020.

[35] T. Steihaug. The conjugate gradient method and trust regions in large scale optimization. *SIAM Journal on NUmerical Analysis*, 20(3):626–637, 1983.

[36] C. Tan, S. Ma, Y. H. Dai, and Y. Qian. Barzilai-Borwein step size for stochastic gradient descent. In *30th Conference on Neural Information Processing Systems (NIPS 2016)*, 2016.

[37] Ph. L. Toint. Towards an efficient sparsity exploiting Newton method for minimization. In I. S. Duff, editor, *Sparse Matrices and Their Uses*, pages 57–88, London, 1981. Academic Press.

[38] B. Wang, H. Zhang, Z. Ma, and W. Chen. Convergence of AdaGrad for non-convex objectives: simple proofs and relaxed assumptions. In *Conference on Learning Theory*, 2023.

[39] L. Wang, H. Wu, and I.A. Matveev. Stochastic gradient method with Barzilai-Borwein step for unconstrained nonlinear optimization. *Journal of Computer and Systems Sciences International*, 60(1):75–86, 2021.

[40] Q. Wang, Ch. Piermarini, Y. Zhu, and F. E. Curtis. Projected stochastic momentum methods for nonlinear equality-constrained optimization for machine learning. *arXiv preprint arXiv:2601.11795*, 2026.

[41] S. Wang, X. Yu, and P. Perdikaris. When and why PINNs fail to train: a neural tangent kernel perspective, 2022.

[42] R. Ward, X. Wu, and L. Bottou. Adagrad stepsizes: sharp convergence over nonconvex landscapes. *Journal of Machine Learning Research*, 21(1):9047–9076, 2020.

[43] T. Zhang and Z. Xu. Convergence of stochastic gradient descent for non-convex problems. In *Proceedings for the COLT Conference*, 2012.

A Appendix. A variant of AS.5

We consider the following assumption in place of AS.5 and show the corresponding complexity result.

Assumption AS.5bis: *There exist the constants $\kappa_{Gg} \geq 0$ and $\kappa_Q \geq 0$ such that*

$$\mathbb{E}_k\left[|(G_k - g_k)^T s_k|\right] \leq \kappa_{Gg}^2 \mathbb{E}_k\left[\|s_k\|^2\right] + \kappa_Q Q_k,$$

with $Q_k > 0$, for all $k \geq 0$.

Lemma A.1 Suppose that AS.3 and AS.5bis hold. Then, for $j \geq 0$,

$$\mathbb{E}_j\left[G_j^T s_j\right] \leq -\frac{\tau\varsigma^2}{2\kappa_B}\,\mathbb{E}_j\left[d_j^T \Delta_j\right] + \kappa_s^2\left(\frac{1}{2}\kappa_B + \kappa_{Gg}^2\right)\mathbb{E}_j\left[\|\Delta_j\|^2\right] + \kappa_Q Q_j, \tag{40}$$

where $\Delta_j \in \mathbb{R}^n$ is the vector whose i-th component is $\Delta_{j,i}$.

Proof. Using AS.5bis and the second part of (10),

$$\begin{aligned}
\mathbb{E}_j\left[G_j^T s_j\right] &= \mathbb{E}_j\left[g_j^T s_j\right] + \mathbb{E}_j\left[(G_j - g_j)^T s_j\right] \\
&\leq \mathbb{E}_j\left[g_j^T s_j\right] + |\mathbb{E}_j\left[(G_j - g_j)^T s_j\right]| \\
&\leq \mathbb{E}_j\left[g_j^T s_j\right] + \kappa_{Gg}^2 |\mathbb{E}_j\left[\|s_j\|^2\right]| + \kappa_Q Q_j \\
&\leq \mathbb{E}_j\left[g_j^T s_j\right] + \kappa_{Gg}^2 \kappa_s^2 \mathbb{E}_j\left[\|\Delta_j\|^2\right] + \kappa_Q Q_j
\end{aligned}$$

and (40) follows from (19). □

Theorem A.2 Suppose that AS.1–AS.4 and AS.5bis hold and that the ADAGB2 algorithm is applied to problem (1). Then

$$\mathbb{E}\left[\underset{j\in\{0,\ldots,k\}}{\text{average}} \|d_j\|\right] \leq \frac{\kappa_{\rm conv}}{\sqrt{k+1}} + \kappa_Q \frac{\sum_{j=0}^k Q_j}{\sqrt{k+1}}, \quad (41)$$

where $\kappa_{\rm conv}$ is given in (23).

Proof. The thesis follows proceeding as in the proof of Theorem 3.4 replacing Γ_0 with $\Gamma_0 + \kappa_Q \sum_{j=0}^k Q_j$. □

If $\kappa_Q > 0$ and $Q_j = \mathcal{O}(1/j)$, we obtain that $\mathbb{E}\left[\text{average}_{j\in\{0,\ldots,k\}} \|d_j\|\right]$ decreases as $\mathcal{O}(\frac{\log(k)}{\sqrt{k+1}})$, while if $Q_j = \mathcal{O}(1/j^p)$ with $p > \frac{1}{2}$, then $\mathbb{E}\left[\text{average}_{j\in\{0,\ldots,k\}} \|d_j\|\right]$ decreases as $\mathcal{O}(\frac{1}{k^{p-1/2}})$. Then, if $p = \frac{3}{4}$ the rate of convergence becomes the standard $\mathcal{O}(\frac{1}{k^4})$.