A unified convergence theory for adaptive first-order methods in the nonconvex case, including AdaNorm, full and diagonal AdaGrad and adaptive Muon

S. Gratton* and Ph. L. Toint†

25 VIII 2026

Abstract

A unified framework for first-order optimization algorithms for nonconvex unconstrained optimization is proposed that uses adaptively preconditioned gradients and includes popular methods such as full and diagonal AdaGrad, AdaNorm, as well an adaptive variant of Muon. This framework (ADPREC) also allows combining heterogeneous geometries across different groups of variables while preserving a unified convergence analysis. A fully stochastic global rate-of-convergence analysis is conducted for all methods in the framework, with and without two types of momentum, using reasonable (although necessarily stronger than bounded variance) assumptions on the gradient oracle and without assuming bounded stochastic gradients or small enough stepsize.

Keywords: Unconstrained nonconvex optimization, first-order methods, global rate of convergence, Adam, AdaGrad, Muon.

1 Introduction

We consider the problem of finding a first-order critical point for the optimization problem

$$\min_X f(X) \tag{1.1}$$

where f is a smooth, possibly nonconvex function of X, whose exact gradient $G(x) = \nabla_X f(X)$ is unknown but can be approached by a random approximation $\widetilde{G}(X, \xi)$ where ξ is an appropriately defined random variable. Stochastic first-order methods for solving problem (1.1) have surged as a major research theme in recent years, essentially motivated by their very successful use in deep learning applications [11]. Because most of these methods do not evaluate the objective function at all[1], they are very robust in the presence of stochastic noise, a crucial feature in these applications. Starting with stochastic gradient descent [43], this domain has seen major advances in the last twenty years, with landmark proposals such as AdaNorm [45, 15, 51], AdaGrad [15, 39], Adam [30], and, more recently, Muon [28]. The number of their variants (see [46, 40, 52, 51, 4, 48, 44] to cite only a few examples) has grown to the point where a comprehensive review is now a major undertaking (and beyond the scope of this paper). Their convergence analysis has followed the same trend, with significant contributions in particular in [58, 42, 52, 34, 14, 23, 18, 29, 21, 17, 36, 49, 26, 27, 57, 33, 35, 41].

*Université de Toulouse, INP, IRIT, Toulouse, France. Email: serge.gratton@enseeiht.fr. Work partially supported by 3IA Artificial and Natural Intelligence Toulouse Institute (ANITI), French "Investing for the Future - PIA3" program under the Grant agreement ANR-19-PI3A-0004"

†Université de Toulouse, INP, IRIT, Toulouse, France, and NAXYS, University of Namur, Namur, Belgium. Email: philippe.toint@unamur.be

[1]They belong to the Objective-Function-Free Optimization (OFFO) class.

A major theme in this domain has been preconditioning, as a cure to the known sensitivity of pure gradient methods to (even modest) problem ill-conditioning. The AdaGrad adaptive diagonal preconditioning strategy proposed in [45, 15] and used in Adam [30], has been at the source of many of the proposals mentioned above, first for unstructured methods (e.g. [52, 21, 1, 2]) and, in the past few years, for methods exploiting the tensor structure of deep-learning applications [48, 35, 44, 56]. The convergence of many of these methods has been investigated, but is often specific to the method considered. From our point of view, notable exceptions are [24] whose interesting interpretation of preconditioning in dual norms covers AdaGrad and some variants in the convex case, [14] covering both Adam and AdaGrad in the nonconvex case, [21] covering AdaGrad and divergent series variants in the nonconvex case, [54, 31] covering AdaGrad, AdaNorm and Shampoo in the convex case, and [41] where the idea of preconditioning in dual norm is applied to discuss (non-adaptive) variants of Muon in the nonconvex case.

As it turns out, results for methods for nonconvex problems involving momentum either establish non-convergence [42, 47][2] or convergence to a mere neighbourhood of a critical point [14], or assume either that gradients are uniformly bounded [42] or (somewhat unrealistically) that the user specifies a stepsize parameter that is sufficiently small compared to the inverse of the problem's Lipschitz constant [23, 26, 53], or to the termination criterion [32, 57].

As far as the authors are aware, the convergence theory for all these proposals (except [54] in the convex case) assumes that the parameters to optimize are of a single type, either a vector of scalars, or a matrix, each having its own dedicated preconditioner. However, practical applications often mix those types (for instance considering together activation levels and biases, and weight matrices in neural networks). Structured second-order methods such as K-FAC [38] approximate curvature information using block-wise Kronecker factorizations, and these methods are typically applied only to selected layers, while simpler and less costly updates are used elsewhere. This practice is for instance recommended in [55, 20, 41, 8], although without analysis for adaptive methods. It is therefore interesting to consider the space of variables as the product space of blocks of parameters of possibly different types. In particular, this has the advantage of allowing a unified view of both vectors and matrix parameters.

Our paper attempts to exploit this observation to propose a unified algorithmic framework, dubbed ADPREC, and to derive an associated general theory. Its main contributions are the following.

1. It provides the first (to the authors' knowledge) truly unified convergence analysis of adaptive preconditioned gradient methods, covering, in the nonconvex case, full and diagonal AdaGrad, AdaNorm and adaptive Muon, with and without momentum, and without assuming boundedness of the gradients or a sufficiently small stepsize parameter.

2. It does so by considering a variable/parameter space structured as the product of blocks of potentially different types (much as in [54]) and by providing new simple proofs based on non-trivial trace inequalities and the theory of operator monotone functions. The ADPREC framework allows combining heterogeneous geometries across different blocks of a Cartesian product space, while preserving a unified and globally consistent complexity analysis. This analysis is conducted under a stochastic assumption on the gradient oracle that allows the coverage of Muon, which, as we show by an example, would be impossible under the standard "bounded variance" condition. This unified analysis is of special interest for adaptive methods because one may fear that the (possibly very) different adaptive stepsizes could generate strong distortions between the rates of convergence across blocks. Besides the obvious difference in proof techniques, it differs from the approach of [54] in that it does not require convexity of the objective function and covers the case where momentum is used.

The paper is organized as follows. Section 2 introduces the structured parameter space and our general algorithmic framework. Its convergence properties are established in Section 3, while

[2] [47] provides a simple example showing that Adam with fixed stepsize might fail to converge in the deterministic setting, irrespective of the choice of its parameters, while [42] gives an example of non-convergence in the stochastic context. Since our present objective is to analyze complexity of (convergent) algorithms, we will not consider Adam any further. We also refer the reader to [60, 14, 59] for more discussion of this issue.

Section 4 details why this framework covers full and diagonal AdaGrad, AdaNorm and adaptive Muon. Some conclusions and perspectives are finally outlined in Section 5.

Notations: In what follows, $\|\cdot\|_E$ denotes the Euclidean norm and the symbols $\succeq$ and $\preceq$ respectively denote the "larger or equal" and "less or equal" relations in the Löwner semi-order on positive semidefinite matrices. If $\|\cdot\|$ is a norm on some space $\mathcal{S}$ endowed with an inner product $\langle\cdot,\cdot\rangle$, its dual norm $\|\cdot\|_*$ is defined by $\|x\|_* = \max_{y\in\mathcal{S}\setminus\{0\}}\langle y, x\rangle/\|y\|$. If x is a vector, $[x]_i$ denotes its i-th component.

2 Parameter space structure and algorithm's abstraction

In order to exploit the possible different types of parameters occurring in problem (1.1), we separate them into L disjoint blocks of internally uniform type (vectors or matrices), that is

$$X = (X_1, \ldots, X_L), \qquad X_\ell = [X]_\ell \in \mathbb{R}^{n_\ell \times m_\ell}.$$

Thus each block contains $d_\ell = n_\ell m_\ell$ parameters the same type, but parameter's types may differ between blocks. The total number of parameters (the dimension of our complete optimization space) is

$$N = \sum_{\ell=1}^{L} n_\ell m_\ell = \sum_{\ell=1}^{L} d_\ell. \tag{2.1}$$

We also define $d_{\max} = \max_{\ell\in\{1,\ldots,L\}} d_\ell$. Each block ℓ is equipped with a norm $\|\cdot\|_\ell$ and its dual norm $\|\cdot\|_{*,\ell}$. The canonical pairing in the Cartesian product space $\mathbb{R}^N = \prod_{\ell=1}^{L}\mathbb{R}^{d_\ell}$ is then defined as

$$\langle U, V\rangle = \sum_{\ell=1}^{L} \langle U_\ell, V_\ell\rangle_F, \quad \text{where} \quad \langle A, B\rangle_F = \mathrm{tr}(A^T B).$$

The (primal) product space is endowed with the norm

$$\|D\|^2 = \sum_{\ell=1}^{L} \|D_\ell\|_\ell^2,$$

whose dual norm is

$$\|Z\|_*^2 = \sum_{\ell=1}^{L} \|Z_\ell\|_{*,\ell}^2. \tag{2.2}$$

(Given a primal norm, the corresponding dual norm is the natural norm to measure gradients or their approximations. The use of dual norms for preconditioning has a long history (e.g. [12, Section 9.4]), and has emerged in the machine learning community following [19, 25, 9, 8, 5] or [41] for instance.)

In this context, our objective is to find a first-order critical point for problem (1.1), that is an X_* such that

$$\|G(X_*)\|_* = 0.$$

In order to define our algorithmic framework to achieve this goal, we choose, for each block ℓ, a measurable normalization operator $\mathcal{N}_\ell(\cdot)$ such that,

$$\mathcal{N}_\ell(D) \begin{cases} \in \arg\max_{\|V\|_\ell\le 1} \langle D, V\rangle_F & \text{if} \quad D \neq 0 \\ = 0 & \text{if} \quad D = 0, \end{cases}$$

which implies that

$$\|\mathcal{N}_\ell(D)\|_\ell = \begin{cases} 1 & \text{if} \quad D \neq 0 \\ 0 & \text{if} \quad D = 0. \end{cases} \qquad \text{and} \quad \langle D, \mathcal{N}_\ell(D)\rangle_F = \|D\|_{*,\ell}. \tag{2.3}$$

(A similar definition was used in [25] for the single-block convex case.) Given that we will study adaptively preconditioned methods, we finally introduce an operator to construct it by successive updates. The operator $\mathcal{U}_{k,\ell}(D) : \mathbb{R}^{n_\ell \times m_\ell} \to S_+^{n_\ell}$ (where $S_+^{n_\ell}$ is the set of positive semidefinite matrices of size n_ℓ) then defines the difference between the preconditioner for block ℓ at iterations $k+1$ and k.

Since computing $G_k = \nabla_X f(X_k)$ is assumed to be either impossible or too expensive, we consider an iterative stochastic approach in which, at iterate X_k, G_k is approximated by an oracle $\widetilde{G}_k$ depending on a random variable ξ_k.

We now motivate these concepts with the simple AdaNorm algorithm [45, 15, 51] on a single block ($L = 1$) in the (self dual) Euclidean norm. This unique block has dimension $n_1 = N$ and $m_1 = 1$, so $\widetilde{G}_k$ and X_k are vectors (now denoted by $\widetilde{g}_k$ and x_k, respectively)

Algorithm 2.1: AdaNorm

Given: a starting point x_0 and constants $\eta, \varsigma > 0$. Set $\gamma_{-1} = \varsigma$.

For $k = 0, 1, \ldots$

1. Draw ξ_k and compute $\widetilde{g}_k = \widetilde{g}(x_k, \xi_k) \approx \nabla_X f(x_k)$,
2. $\gamma_k = \gamma_{k-1} + \|\widetilde{g}_k\|^2$,
3. $x_{k+1} = x_k - \eta \, \dfrac{\widetilde{g}_k}{\sqrt{\gamma_k}}$.

For our purpose, we rewrite the update to γ_k, now in matrix form, as

$$\Gamma_k = \Gamma_{k-1} + \|\widetilde{G}_k\|^2 I = \Gamma_{k-1} + \mathcal{U}_{k,1}(\widetilde{G}_k),$$

defining the operator $\mathcal{U}_{k,1}(D) = \|D\|^2 I$ for AdaNorm. The iterate update may also be reformulated, in its matrix form, as

$$X_{k+1} = X_k - \eta \, \frac{\|\widetilde{G}_k\|}{\sqrt{\Gamma_k}} \cdot \frac{\widetilde{G}_k}{\|\widetilde{G}_k\|} = X_k - \eta \, \|Z_k\| \cdot \mathcal{N}_1(Z_k)$$

where

$$Z_k = \frac{\widetilde{G}_k}{\sqrt{\gamma_k}} = \Gamma_k^{-1/2} \widetilde{G}_k \quad \text{and} \quad \mathcal{N}_1(Z_k) = \frac{\widetilde{G}_k}{\|\widetilde{G}_k\|} = \frac{Z_k}{\|Z_k\|},$$

and $\mathcal{N}_1(\cdot)$ is the normalization operator in the Euclidean norm. Thus the familiar AdaNorm may be reformulated using the $\mathcal{U}_{k,1}(\cdot)$ and $\mathcal{N}_1(\cdot)$ operators. We claim (and verify in Section 4) that this is not only the case for AdaNorm, but the same formalism also applies to other methods of interest.

Having motivated the introduction of the $\mathcal{N}_\ell(\cdot)$ and $\mathcal{U}_{k,\ell}(\cdot)$, we now state our simple but general algorithmic ADPREC framework on the following page.

We now illustrate these concepts by showing that familiar algorithms fit in the ADPREC framework. For clarity of exposition, we use lower case symbols for vectors.

2.1 AdaNorm [45, 15, 51]

Taking the AdaNorm update on a given block ℓ corresponds to choosing an isotropic scalar geometry in the relevant subspace. Reformulating the above description for a general block, we see that, if block ℓ has dimension $n_\ell \times 1$,

$$\Gamma_{-1,\ell} = \varsigma I_{n_\ell} \quad \text{and} \quad \Gamma_{k,\ell} = \Gamma_{k-1,\ell} + \frac{\|\widetilde{g}_{k,\ell}\|_\ell^2}{n_\ell} I_{n_\ell} \quad (k \geq 0),$$

Algorithm 2.2: ADPREC

Given: a starting point X_0, operators $\{\mathcal{U}_{k,\ell}\}_{k\geq 0,\ell\in\{1,\ldots,L\}}$ and $\{\mathcal{N}_\ell\}_{\ell\in\{1,\ldots,L\}}$ and constants $\eta,\varsigma>0$. Set $\Gamma_{-1,\ell}=\varsigma I_\ell$ for $\ell\in\{1,\ldots,L\}$.

For $k=0,1,\ldots$

1. Draw ξ_k and compute $\widetilde{G}_k=\widetilde{G}(X_k,\xi_k)=(\widetilde{G}_{k,1},\ldots,\widetilde{G}_{k,L})\approx G(X_k)$, [gradient approx.]
2. On each block $\ell\in\{1,\ldots,L\}$, compute

$$\begin{aligned}\Gamma_{k,\ell} &= \Gamma_{k-1,\ell}+\mathcal{U}_{k,\ell}(\widetilde{G}_{k,\ell}), && \text{[preconditioner update]} \quad (2.4)\\ Z_{k,\ell} &= \Gamma_{k,\ell}^{-1/2}\widetilde{G}_{k,\ell}, && \text{[preconditioned gradient]} \quad (2.5)\\ X_{k+1,\ell} &= X_{k,\ell}-\eta\,\|Z_{k,\ell}\|_{*,\ell}\,\mathcal{N}_\ell(Z_{k,\ell}). && \text{[scaled normalized step]} \quad (2.6)\end{aligned}$$

($\widetilde{g}_{k,\ell}$ is now a vector) so that the contribution of this block to the step is

$$s_{k,\ell}=-\eta\,\Gamma_{k,\ell}^{-1/2}\widetilde{g}_{k,\ell}=-\frac{\eta}{\sqrt{\gamma_{k,\ell}}}\,\widetilde{g}_{k,\ell}=-\eta\frac{\widetilde{g}_{k,\ell}}{\sqrt{\gamma_{k,\ell}}}\frac{\widetilde{g}_{k,\ell}}{\|\widetilde{g}_{k,\ell}\|}$$

when $\Gamma_{k,\ell}=\gamma_{k,\ell}I_{n_\ell}$. This is exactly the AdaNorm update on block ℓ. In ADPREC, this therefore corresponds to choosing

$$m_\ell=1,\qquad \|\cdot\|_{*,\ell}=\|\cdot\|_\ell=\|\cdot\|_E,\qquad \mathcal{N}_\ell(\widetilde{G}_{k,\ell})=\frac{\widetilde{G}_{k,\ell}}{\|\widetilde{G}_{k,\ell}\|_\ell},\quad\text{and}\quad \mathcal{U}_{k,\ell}(\widetilde{G}_{k,\ell})=\frac{\|\widetilde{G}_{k,\ell}\|_\ell^2}{n_\ell}I_{n_\ell}. \tag{2.7}$$

2.2 Full AdaGrad [15]

Taking the "full" version of AdaGrad [15] update on block ℓ corresponds to choosing, on a given block ℓ, a full matrix-valued geometry. More precisely, if block ℓ has dimension n_ℓ, one considers

$$\Gamma_{-1,\ell}=\varsigma I_{n_\ell}\quad\text{and}\quad \Gamma_{k,\ell}=\Gamma_{k-1,\ell}+\widetilde{g}_{k,\ell}\widetilde{g}_{k,\ell}^T\quad(k\geq 0),$$

so that the contribution of this block to the step is $s_{k,\ell}=-\eta\,\Gamma_{k,\ell}^{-1/2}\widetilde{g}_{k,\ell}$. This is exactly the full AdaGrad update on block ℓ. In the ADPREC framework, this corresponds to choosing

$$m_\ell=1,\qquad \|\cdot\|_{*,\ell}=\|\cdot\|_\ell=\|\cdot\|_E,\qquad \mathcal{N}_\ell(\widetilde{G}_{k,\ell})=\widetilde{G}_{k,\ell}/\|\widetilde{G}_{k,\ell}\|_E,\quad\text{and}\quad \mathcal{U}_{k,\ell}(\widetilde{G}_{k,\ell})=\widetilde{G}_{k,\ell}\widetilde{G}_{k,\ell}^T. \tag{2.8}$$

2.3 Diagonal AdaGrad [15, 39]

We now turn to the use of the diagonal/component-wise AdaGrad update on a given block ℓ. If block ℓ has dimension $n_\ell\times 1$, this geometry is defined by

$$\Gamma_{-1,\ell}=\varsigma I_{n_\ell}\quad\text{and}\quad \Gamma_{k,\ell}=\Gamma_{k-1,\ell}+\operatorname{diag}\big([\widetilde{g}_{k,\ell}]_i^2\big)\quad(k\geq 0)\quad\text{and}\quad x_{k+1,\ell}=x_{k,\ell}-\eta\Gamma_{k,\ell}^{-1/2}\widetilde{g}_{k,\ell}.$$

In our framework, this corresponds to choosing

$$m_\ell=1,\qquad \|\cdot\|_{*,\ell}=\|\cdot\|_\ell=\|\cdot\|_E,\qquad \mathcal{U}_{k,\ell}(\widetilde{g}_{k,\ell})=\operatorname{diag}\big([\widetilde{g}_{k,\ell}]_i^2\big),\quad\text{and}\quad \mathcal{N}_\ell(\widetilde{g}_{k,\ell})=\frac{\widetilde{g}_{k,\ell}}{\|\widetilde{g}_{k,\ell}\|_E}, \tag{2.9}$$

Observe that we could also view the diagonal AdaGrad algorithm as a Cartesian product of scalar AdaNorm. Also note that, as discussed in Section 3.3, momentum in the form (3.48)-(3.50) or (3.63)-(3.65) may be introduced on this block, resulting in an ADPREC variant somewhat similar to Adam [30]. It is theoretically interesting that this variant enjoys the convergence behaviour described by Theorem 3.13 and Corollary 3.11.

2.4 Adaptive MUON/AdaGO [28, 44, 56]

We now consider an AdaGrad-inspired adaptive version of MUON [28], described as AdaGO for the single block case in [56]. In our framework, this corresponds to assigning, for each block ℓ, a geometry adapted to the structure of neural network parameters. We distinguish two types of blocks. For $\ell \in \mathcal{M}$, $G_{k,\ell}$ is a matrix of size $n_\ell \times m_\ell$, while for $\ell \in \mathcal{A} = \{1, \ldots, L\} \setminus \mathcal{M}$, $G_{k,\ell}$ is a vector in $\mathbb{R}^{n_\ell}$. For $\ell \in \mathcal{A}$, we use

$$\|\cdot\|_\ell = \|\cdot\|_E = \|\cdot\|_{*,\ell},$$

while for $\ell \in \mathcal{M}$,

$$\|\cdot\|_\ell = \|\cdot\|_s, \qquad \|\cdot\|_{*,\ell} = \|\cdot\|_*,$$

where $\|\cdot\|_s$ is the spectral norm and $\|\cdot\|_*$ the nuclear norm. The MUON normalization is defined by

$$\mathcal{N}_\ell(\widetilde{G}_{k,\ell}) = \begin{cases} U_{k,\ell}V_{k,\ell}^T & \text{if} \quad \ell \in \mathcal{M} \ \text{ and } \ \widetilde{G}_{k,\ell} = U_{k,\ell}\Sigma_{k,\ell}V_{k,\ell}^T, \\ \widetilde{G}_{k,\ell}/\|\widetilde{G}_{k,\ell}\|_E & \text{if} \quad \ell \in \mathcal{A}. \end{cases} \tag{2.10}$$

We then define, for each block ℓ,

$$\Gamma_{k,\ell} = \gamma_{k,\ell}\, I_{d_\ell}, \quad \text{where} \quad \gamma_{-1,\ell} = \varsigma \quad \text{and} \quad \gamma_{k,\ell} = \begin{cases} \gamma_{k-1,\ell} + \|\widetilde{G}_{k,\ell}\|_*^2/d_\ell & \text{if} \quad \ell \in \mathcal{M}, \\ \gamma_{k-1,\ell} + \|\widetilde{G}_{k,\ell}\|_E^2/d_\ell & \text{if} \quad \ell \in \mathcal{A}. \end{cases} \tag{2.11}$$

This corresponds to choosing, for each block ℓ,

$$\mathcal{U}_{k,\ell}(\widetilde{G}_{k,\ell}) = \begin{cases} \frac{\|\widetilde{G}_{k,\ell}\|_*^2}{d_\ell} I_{d_\ell} & \text{if} \quad \ell \in \mathcal{M}, \\ \frac{\|\widetilde{G}_{k,\ell}\|_E^2}{d_\ell} I_{d_\ell} & \text{if} \quad \ell \in \mathcal{A}. \end{cases}$$

We also have that $Z_{k,\ell} = \frac{1}{\sqrt{\gamma_{k,\ell}}}\widetilde{G}_{k,\ell}$. With these choices, the resulting ADPREC algorithm corresponds to a block-wise adaptive MUON method, with stepsize $\eta_{\text{muon},\ell} = \eta/\sqrt{d_\ell}$.

2.5 A few comments on ADPREC

At each iteration, the random variable ξ_k (whose distribution may depend on X_k) is defined on some probability space $(\Omega, \mathcal{F}, \mathbb{P})$ ($\mathbb{E}$ denotes the corresponding expectation). The expectation conditioned to knowing $(\xi_0, \ldots, \xi_{k-1})$ is denoted by the symbol $\mathbb{E}_k[\cdot]$.

For each choice of $\mathcal{U}_{k,\ell}$ and $\mathcal{N}_\ell$, the resulting algorithm thus defines a random process

$$(\widetilde{G}_k, Z_k, \Gamma_k, X_k),$$

where $\widetilde{G}_k = (\widetilde{G}_{k,1}, \ldots, \widetilde{G}_{k,L})$, $Z_k = (Z_{k,1}, \ldots, Z_{k,L})$, $\Gamma_k = (\Gamma_{k,1}, \ldots, \Gamma_{k,L})$ and $X_k = (X_{k,1}, \ldots, X_{k,L})$. At first sight, it may seem that the algorithm performs separate optimization on each block, but this view is deceptive because interaction between the blocks occurs in the computation of the approximate gradient at the full-dimensional iterate X_k, therefore requiring synchronization between blocks.

The name ADPREC alludes to the fact that $\Gamma_{k,\ell}$ may be interpreted as an adaptive preconditioner for block ℓ and (2.5) therefore produces an adaptively preconditioned (approximate) gradient $Z_{k,\ell}$.

Note that we allow changing the update formula in (2.4) for $\Gamma_{k,\ell}$ at each iteration ($\mathcal{U}_{k,\ell}$ could depend on k), although, as far as the authors know, this is not done in practice. We could also consider making the stepsize η and/or the initial preconditioner size ς dependent on the block (that is using η_ℓ and/or ς_ℓ), at the cost of introducing more constants in our already intricate analysis.

3 Convergence analysis

The convergence analysis for ADPREC hinges on the following standard assumptions.

Assumption 1 (Boundedness) *There exists a constant $f_{\rm low}$ such that $f(X) \geq f_{\rm low}$ for all X.*

Assumption 2 (Smoothness) *The objective function f is continuously differentiable and has a Lipschitz continuous gradient, that is there exists a constant $L_G \geq 0$ such that, for all $X, Y \in \mathbb{R}^N$, $\|G(X) - G(Y)\|_* \leq L_G \|X - Y\|$.*

We now introduce two conditions that define the class of algorithms within the ADPREC framework, for which we develop our theory. We present them as assumptions *on the algorithm(s)*, rather than on the problem. Although seemingly abstract, we will demonstrate in Section 4 that they do hold for the methods introduced in Sections 2.1 to 2.4.

Assumption 3 (Structural identities) *For all $k \geq 0$ and all $\ell \in \{1, \ldots, L\}$, we have that*

$$\|Z_{k,\ell}\|_{*,\ell} \left\langle \widetilde{G}_{k,\ell}, \mathcal{N}_\ell(Z_{k,\ell}) \right\rangle_F = \mathrm{tr}\big(\Gamma_{k,\ell}^{-1/2} \mathcal{U}_{k,\ell}(\widetilde{G}_{k,\ell})\big), \tag{3.1}$$

and

$$\|Z_{k,\ell}\|_{*,\ell}^2 = \mathrm{tr}\big(\Gamma_{k,\ell}^{-1} \mathcal{U}_{k,\ell}(\widetilde{G}_{k,\ell})\big). \tag{3.2}$$

As will become clear in Lemma 3.1 just below, these identities reformulate the terms associated with first-order descent (for (3.1)) and second-order perturbations (for (3.2)) in terms of traces. Note that (3.1) and (3.2) are "iteration-specific" conditions, in that they only involve $\Gamma_{k,\ell}$ and are independent of the value of $\Gamma_{k-1,\ell}$.

Assumption 4 (Gradient-preconditioner compatibility) *There exists a constant $\kappa_\circ > 0$ such that, for each block $\ell \in \{1, \ldots, L\}$, each $k \geq 0$ and all G_ℓ,*

$$\|\widetilde{G}_\ell\|_{*,\ell}^2 \leq \kappa_\circ^2 \ \mathrm{tr}\left(\mathcal{U}_{k,\ell}(\widetilde{G}_\ell)\right). \tag{3.3}$$

Because $\mathcal{U}_{k,\ell}(\cdot)$ is used to build the preconditioners $\Gamma_{k,\ell}$, this last assumption establishes the link between them and the measures of (approximate) optimality $\widetilde{G}_{k,\ell}$. (Note that we could choose $\kappa_\circ$ depending on ℓ, but we use a single constant for simplicity.)

We of course need to be more specific on what we assume on the oracle $\widetilde{G}_k$, but postpone the precise nature of our requirements to the statements of the results where they are needed.

Our first step is to analyze the (expected) descent at iteration k.

Lemma 3.1 Suppose that Assumption 2 holds. Then

$$\begin{aligned}\mathbb{E}_k[f(X_{k+1})] \leq & f(X_k) - \eta \sum_{\ell=1}^{L} \mathbb{E}_k\left[\|Z_{k,\ell}\|_{*,\ell} \left\langle \widetilde{G}_{k,\ell}, \mathcal{N}_\ell(Z_{k,\ell}) \right\rangle_F\right] \\ & + \eta \sum_{\ell=1}^{L} \mathbb{E}_k\left[\|Z_{k,\ell}\|_{*,\ell} \left| \left\langle \widetilde{G}_{k,\ell} - G_{k,\ell}, \mathcal{N}_\ell(Z_{k,\ell}) \right\rangle_F \right|\right] + \frac{L_G \eta^2}{2} \sum_{\ell=1}^{L} \mathbb{E}_k\left[\|Z_{k,\ell}\|_{*,\ell}^2\right].\end{aligned} \tag{3.4}$$

Proof. By Lipschitz continuity of the gradient (Assumption 2),

$$f(X_{k+1}) \leq f(X_k) + \langle G_k, X_{k+1} - X_k \rangle + \frac{L_G}{2} \|X_{k+1} - X_k\|^2.$$

Thus, using (2.5) and (2.6), we obtain

$$f(X_{k+1}) \leq f(X_k) - \eta \sum_{\ell=1}^{L} \|Z_{k,\ell}\|_{*,\ell} \langle G_{k,\ell}, \mathcal{N}_\ell(Z_{k,\ell}) \rangle_F + \frac{L_G \eta^2}{2} \sum_{\ell=1}^{L} \|Z_{k,\ell}\|_{*,\ell}^2 \|\mathcal{N}_\ell(Z_{k,\ell})\|_\ell^2.$$

Taking conditional expectations, inserting $\widetilde{G}_{k,\ell}$ and using (2.3) yields (3.4). □

3.1 Trace inequalities

Our argument now takes a little linear algebra detour for proving two useful inequalities about the trace of matrices of interest. We first derive an inequality involving the trace of the inverse square root of the sum of two matrices.

Lemma 3.2 Let A be a positive definite matrix and B be a positive semidefinite matrix. Then

$$\mathrm{tr}\big((A+B)^{-1/2}B\big) \geq \mathrm{tr}((A+B)^{1/2}) - \mathrm{tr}(A^{1/2}). \tag{3.5}$$

Proof. Set $X = (A+B)^{1/2}$ and $Y = A^{1/2}$. Clearly, $X \succ 0$, $Y \succ 0$, and $X + Y \succ 0$ and is therefore invertible. Moreover

$$B = X^2 - Y^2 = X(X-Y) + (X-Y)Y. \tag{3.6}$$

Multiplying this inequality on the right by $(X+Y)^{-1}$ gives

$$B(X+Y)^{-1} = X(X-Y)(X+Y)^{-1} + (X-Y)Y(X+Y)^{-1}.$$

Taking the trace and using cyclicity, we obtain that

$$\begin{aligned}
\mathrm{tr}\big(B(X+Y)^{-1}\big) &= \mathrm{tr}\big((X+Y)^{-1}X(X-Y)\big) + \mathrm{tr}\big((X+Y)^{-1}(X-Y)Y\big) \\
&= \mathrm{tr}\Big((X+Y)^{-1}\big(X(X-Y) + (X-Y)Y\big)\Big) \\
&= \mathrm{tr}\Big((X+Y)^{-1}\big((X-Y)(X+Y) - [X,Y]\big)\Big)
\end{aligned}$$

where $[X,Y] = XY - YX$ is the commutator between X and Y. Since X and Y are Hermitian, $[X,Y]$ is skew-Hermitian and its trace is zero. Hence, using cyclicity (twice),

$$\begin{aligned}
\mathrm{tr}\big((X+Y)^{-1}B\big) = \mathrm{tr}\big(B(X+Y)^{-1}\big) &= \mathrm{tr}\big((X+Y)^{-1}(X-Y)(X+Y)\big) - \mathrm{tr}\big((X+Y)^{-1}[X,Y]\big) \\
&= \mathrm{tr}\big((X+Y)^{-1}(X-Y)(X+Y)\big) \\
&= \mathrm{tr}(X-Y)
\end{aligned}$$

Now, since $Y \succeq 0$, we have that $X + Y \succeq X$. Since both X and $X+Y$ are positive definite and since inversion reverses the Löwner semi-order, we deduce that

$$(X+Y)^{-1} \preceq X^{-1} = (A+B)^{-1/2}.$$

and thus $(A+B)^{-1/2} - (X+Y)^{-1} \succeq 0$. But, since B is symmetric positive semidefinite, we deduce, again using cyclicity, that

$$\mathrm{tr}\Big(\big((A+B)^{-1/2} - (X+Y)^{-1}\big)B\Big) = \mathrm{tr}\Big(B^{1/2}\big((A+B)^{-1/2} - (X+Y)^{-1}\big)B^{1/2}\Big) \succeq 0,$$

and therefore that

$$\mathrm{tr}\Big((A+B)^{1/2}\Big) - \mathrm{tr}\Big(A^{1/2}\Big) = \mathrm{tr}(X-Y) = \mathrm{tr}\Big((X+Y)^{-1}B\Big) \leq \mathrm{tr}\Big((A+B)^{-1/2}B\Big),$$

which is (3.5). □

The next step of our argument uses the Löwner semi-order $\succeq$ on symmetric positive-semidefinite matrices and the theory of operator monotone functions (see [10, Chapter V], for instance).

Lemma 3.3 Let A be a positive definite matrix and B be a positive semidefinite matrix, and let $\phi : (0,\infty) \to \mathbb{R}$ be a C^1 function. Suppose that ϕ' is operator monotone decreasing on $(0,\infty)$, meaning that for any positive definite matrices X, Y, $X \preceq Y$ implies that $\phi'(X) \succeq \phi'(Y)$. Then

$$\mathrm{tr}\big(\phi'(A+B)B\big) \leq \mathrm{tr}\big(\phi(A+B)-\phi(A)\big) \leq \mathrm{tr}\big(\phi'(A)B\big). \tag{3.7}$$

Proof. Define $g(t) \stackrel{\rm def}{=} \mathrm{tr}\big(\phi(A+tB)\big)$ for $t \in [0,1]$. Using the standard derivative formula for spectral functions under the trace, we see that

$$g'(t) = \mathrm{tr}\big(\phi'(A+tB)B\big). \tag{3.8}$$

Since $B \succeq 0$, we also have that

$$A \preceq A+tB \preceq A+B \qquad (t \in [0,1]). \tag{3.9}$$

and hence, because ϕ' is operator monotone decreasing, that

$$\phi'(A+B) \preceq \phi'(A+tB) \preceq \phi'(A).$$

Now if $M \preceq N$ and $B \succeq 0$, then $\mathrm{tr}((N-M)B) = \mathrm{tr}(B^{1/2}(N-M)B^{1/2}) \geq 0$, and thus $\mathrm{tr}(MB) \leq \mathrm{tr}(NB)$. Using this inequality in (3.9) and the identity (3.8), we therefore obtain that

$$\mathrm{tr}\big(\phi'(A+B)B\big) \leq g'(t) \leq \mathrm{tr}\big(\phi'(A)B\big).$$

Integrating over $t \in [0,1]$ then yields (3.7). □

We now use the above and a classical result by K. Löwner to deduce two helpful inequalities on the trace of matrices generated by the ADPREC algorithm.

Lemma 3.4 We have that, for all $k \geq 0$,

$$\sum_{\ell=1}^{L} \mathrm{tr}(\Gamma_{k,\ell}^{1/2}) - \sum_{\ell=1}^{L} \mathrm{tr}(\Gamma_{-1,\ell}^{1/2}) \leq \sum_{j=0}^{k} \sum_{\ell=1}^{L} \mathrm{tr}\Big(\Gamma_{j,\ell}^{-1/2} \mathcal{U}_{j,\ell}(\widetilde{G}_{j,\ell})\Big) \tag{3.10}$$

and

$$\sum_{j=0}^{k} \sum_{\ell=1}^{L} \mathrm{tr}\Big(\Gamma_{j,\ell}^{-1} \mathcal{U}_{j,\ell}(\widetilde{G}_{j,\ell})\Big) \leq \sum_{\ell=1}^{L} \mathrm{tr}\Big(\log(\Gamma_{k,\ell})\Big) - \sum_{\ell=1}^{L} \mathrm{tr}\Big(\log(\Gamma_{-1,\ell})\Big). \tag{3.11}$$

Proof. Let

$$A = \Gamma_{j-1,\ell} \quad \text{and} \quad B = \mathcal{U}_{j,\ell}(\widetilde{G}_{j,\ell}). \tag{3.12}$$

Then $\Gamma_{j,\ell} = A + B$ and Theorem 3.2 gives that, for $j \in \{0,\ldots,k\}$ and $\ell \in \{1,\ldots,L\}$,

$$\mathrm{tr}(\Gamma_{j,\ell}^{1/2}) - \mathrm{tr}(\Gamma_{j-1,\ell}^{1/2}) \leq \mathrm{tr}\Big(\Gamma_{j,\ell}^{-1/2} \mathcal{U}_{j,\ell}(\widetilde{G}_{j,\ell})\Big).$$

Summing this inequality for $j \in \{0,\ldots,k\}$ and $\ell \in \{1,\ldots,L\}$ yields (3.10). In order to prove (3.11), let $\phi(t) = \log t$. Then $\phi'(t) = \frac{1}{t}$. which is operator monotone decreasing (see [37] as cited in [10, p. 149]). Applying the first inequality of (3.7) in Theorem 3.3 then yields that

$$\mathrm{tr}\Big((A+B)^{-1}B\Big) \leq \mathrm{tr}\Big(\log(A+B) - \log A\Big),$$

which, with (3.12), ensures that, for $j \in \{0, \ldots, k\}$ and $\ell \in \{1, \ldots, L\}$,

$$\mathrm{tr}\Big(\Gamma_{j,\ell}^{-1}\mathcal{U}_{j,\ell}(\widetilde{G}_{j,\ell})\Big) \leq \mathrm{tr}\Big(\log(\Gamma_{j,\ell})\Big) - \mathrm{tr}\Big(\log(\Gamma_{j-1,\ell})\Big).$$

Summing this inequality for $j \in \{0, \ldots, k\}$ and $\ell \in \{1, \ldots, L\}$ now yields (3.11). □

3.2 Telescoping inequality and global rate of convergence

We next introduce a variant of the classical telescoping argument, leading to the central inequality in our analysis.

Theorem 3.5 Suppose that Assumption 1 and 3 hold. Suppose also that, for some $\omega \geq 0$ and $\nu_k \geq 0$

$$\sum_{j=0}^{k} \mathbb{E}\Big[\|\widetilde{G}_j - G_j\|_*^2\Big] \leq \nu_k^2 + \omega^2 \sum_{j=0}^{k}\sum_{\ell=1}^{L} \mathbb{E}\big[\|Z_{j,\ell}\|_{*,\ell}^2\big]. \tag{3.13}$$

Define

$$\Delta_k \stackrel{\rm def}{=} \mathbb{E}\left[\sum_{\ell=1}^{L} \mathrm{tr}(\log \Gamma_{k,\ell}) - \sum_{\ell=1}^{L} \mathrm{tr}(\log \Gamma_{-1,\ell})\right]. \tag{3.14}$$

Then, for every $k \geq 0$,

$$\mathbb{E}\left[\sum_{\ell=1}^{L} \mathrm{tr}(\Gamma_{k,\ell}^{1/2})\right] \leq \kappa_{\rm gap} + \nu_k \sqrt{\Delta_k} + \left(\omega + \frac{L_G\eta}{2}\right) \Delta_k, \tag{3.15}$$

where

$$\kappa_{\rm gap} \stackrel{\rm def}{=} \frac{f(X_0) - f_{\rm low}}{\eta} + \sqrt{\varsigma}\, N. \tag{3.16}$$

Proof. Taking the full expectation in the conditional descent inequality (3.4) and summing for $j = 0$ to k gives

$$\begin{aligned}\eta \sum_{j=0}^{k}\sum_{\ell=1}^{L} \mathbb{E}\Big[\|Z_{j,\ell}\|_{*,\ell} \left\langle \widetilde{G}_{j,\ell}, \mathcal{N}_\ell(Z_{j,\ell})\right\rangle_F\Big] &\leq f(X_0) - f_{\rm low} \\ &+ \eta \sum_{j=0}^{k}\sum_{\ell=1}^{L} \mathbb{E}\Big[\|Z_{j,\ell}\|_{*,\ell} \left|\left\langle \widetilde{G}_{j,\ell} - G_{j,\ell}, \mathcal{N}_\ell(Z_{j,\ell})\right\rangle_F\right|\Big] + \frac{L_G\eta^2}{2} \sum_{j=0}^{k}\sum_{\ell=1}^{L} \mathbb{E}\big[\|Z_{j,\ell}\|_{*,\ell}^2\big].\end{aligned} \tag{3.17}$$

Using successively (3.1) and (3.10), we obtain that

$$\begin{aligned}\eta \sum_{j=0}^{k}\sum_{\ell=1}^{L} \mathbb{E}\Big[\|Z_{j,\ell}\|_{*,\ell} \left\langle \widetilde{G}_{j,\ell}, \mathcal{N}_\ell(Z_{j,\ell})\right\rangle_F\Big] &= \eta \sum_{j=0}^{k}\sum_{\ell=1}^{L} \mathbb{E}\Big[\mathrm{tr}\big(\Gamma_{j,\ell}^{-1/2}\mathcal{U}_{j,\ell}(\widetilde{G}_{j,\ell})\big)\Big] \\ &\geq \eta\, \mathbb{E}\left[\sum_{\ell=1}^{L} \mathrm{tr}(\Gamma_{k,\ell}^{1/2}) - \sum_{\ell=1}^{L} \mathrm{tr}(\Gamma_{-1,\ell}^{1/2})\right].\end{aligned} \tag{3.18}$$

Now observe that the definition of the dual norm and (2.3) give that

$$\left|\left\langle \widetilde{G}_{j,\ell} - G_{j,\ell}, \mathcal{N}_\ell(Z_{j,\ell})\right\rangle_F\right| \leq \|\widetilde{G}_{j,\ell} - G_{j,\ell}\|_{*,\ell}\, \|\mathcal{N}_\ell(Z_{j,\ell})\|_\ell \leq \|\widetilde{G}_{j,\ell} - G_{j,\ell}\|_{*,\ell}.$$

Therefore

$$\mathbb{E}\Big[\|Z_{j,\ell}\|_{*,\ell}\,|\,\Big\langle \widetilde{G}_{j,\ell}-G_{j,\ell},\mathcal{N}_\ell(Z_{j,\ell})\Big\rangle_F\,|\Big] \le \sqrt{\mathbb{E}\Big[\|\widetilde{G}_{j,\ell}-G_{j,\ell}\|_{*,\ell}^2\Big]}\,\sqrt{\mathbb{E}\Big[\|Z_{j,\ell}\|_{*,\ell}^2\Big]}.$$

Observe now that the Cauchy-Schwartz inequality also implies that, for any vectors a and b with nonnegative components,

$$\sum_j \sqrt{a_j}\sqrt{b_j} \le \|\sqrt{a}\|_E\,\|\sqrt{b}\|_E = \sqrt{\sum_j a_j}\sqrt{\sum_j b_j}. \tag{3.19}$$

Hence, summing over ℓ and using (2.2), we obtain that

$$\sum_{\ell=1}^L \mathbb{E}\Big[\|Z_{j,\ell}\|_{*,\ell}\,|\,\Big\langle \widetilde{G}_{j,\ell}-G_{j,\ell},\mathcal{N}_\ell(Z_{j,\ell})\Big\rangle_F\,|\Big] \le \sqrt{\mathbb{E}\Big[\|\widetilde{G}_j-G_j\|_*^2\Big]}\,\sqrt{\sum_{\ell=1}^L \mathbb{E}\Big[\|Z_{j,\ell}\|_{*,\ell}^2\Big]}.$$

Summing now over j, using $\sqrt{a+b}\le\sqrt{a}+\sqrt{b}$ and (3.19) again, we deduce that

$$\begin{aligned}
&\eta\sum_{j=0}^k\sum_{\ell=1}^L \mathbb{E}\Big[\|Z_{j,\ell}\|_{*,\ell}\,|\,\Big\langle \widetilde{G}_{j,\ell}-G_{j,\ell},\mathcal{N}_\ell(Z_{j,\ell})\Big\rangle_F\,|\Big]\\
&\qquad\le \eta\sqrt{\sum_{j=0}^k \mathbb{E}\Big[\|\widetilde{G}_j-G_j\|_*^2\Big]}\,\sqrt{\sum_{j=0}^k\sum_{\ell=1}^L \mathbb{E}\Big[\|Z_{j,\ell}\|_{*,\ell}^2\Big]}\\
&\qquad\le \eta\nu_k\sqrt{\sum_{j=0}^k\sum_{\ell=1}^L \mathbb{E}\Big[\|Z_{j,\ell}\|_{*,\ell}^2\Big]} + \eta\omega\sum_{j=0}^k\sum_{\ell=1}^L \mathbb{E}\big[\|Z_{j,\ell}\|_{*,\ell}^2\big]
\end{aligned}$$

where we used (3.13) to obtain the last inequality. But, by (3.2), (3.11) and (3.14),

$$\sum_{j=0}^k\sum_{\ell=1}^L \mathbb{E}\big[\|Z_{j,\ell}\|_{*,\ell}^2\big] = \sum_{j=0}^k\sum_{\ell=1}^L \mathbb{E}\Big[\mathrm{tr}\big(\Gamma_{j,\ell}^{-1}\mathcal{U}_{j,\ell}(\widetilde{G}_{j,\ell})\big)\Big] \le \Delta_k, \tag{3.20}$$

so that

$$\eta\sum_{j=0}^k\sum_{\ell=1}^L \mathbb{E}\Big[\|Z_{j,\ell}\|_{*,\ell}\,|\,\Big\langle \widetilde{G}_{j,\ell}-G_{j,\ell},\mathcal{N}_\ell(Z_{j,\ell})\Big\rangle_F\,|\Big] \le \eta\,\nu_k\,\sqrt{\Delta_k}+\eta\omega\Delta_k. \tag{3.21}$$

Similarly, using (2.3) and (3.20), the quadratic term satisfies

$$\frac{L_G\eta^2}{2}\sum_{j=0}^k\sum_{\ell=1}^L \mathbb{E}\big[\|Z_{j,\ell}\|_{*,\ell}^2\|\mathcal{N}_\ell(Z_{j,\ell})\|_\ell^2\big] = \frac{L_G\eta^2}{2}\sum_{j=0}^k\sum_{\ell=1}^L \mathbb{E}\big[\|Z_{j,\ell}\|_{*,\ell}^2\big] \le \frac{L_G\eta^2}{2}\Delta_k. \tag{3.22}$$

Substituting (3.18), (3.21) and (3.22) into (3.17) then yields that

$$\eta\,\mathbb{E}\left[\sum_{\ell=1}^L \mathrm{tr}(\Gamma_{k,\ell}^{1/2})-\sum_{\ell=1}^L \mathrm{tr}(\Gamma_{-1,\ell}^{1/2})\right] \le f(X_0)-f_{\rm low}+\eta\,\nu_k\,\sqrt{\Delta_k}+\left(\eta\omega+\frac{L_G\eta^2}{2}\right)\Delta_k, \tag{3.23}$$

which is exactly (3.15)-(3.16) after taking into account that $\Gamma_{-1,\ell}=\varsigma I_\ell$ for all $\ell\in\{1,\ldots,L\}$. □

Observe that the requirement (3.13) defines an upper bound on the *cumulative* total variance of the gradient oracle over all past iterates, an approach more general than assuming conditional variance at every iteration. In particular, it allows large variance at early iterations provided later iterations

compensate. Trade-offs between different realizations are also theoretically possible. Also note that, since we will prove that Γ_k grows very slowly (see (3.28) below), and Z_k is the preconditioned approximate gradient, condition (3.13) is akin to a cumulative affine variance condition of the form

$$\sum_{j=0}^{k}\sum_{\ell=1}^{L}\mathbb{E}_j\left[\|\widetilde{G}_{j,\ell}-G_{j,\ell}\|_{*,\ell}^2\right] \le \nu_k^2+\omega^2\sum_{j=0}^{k}\sum_{\ell=1}^{L}\mathbb{E}_j\left[\|G_{j,\ell}\|_{*,\ell}^2\right],$$

whose iteration-wise variant has already been used in analysis of first-order methods (see [18, 49] or, for the stronger " affine-* " version, [4]). In particular, the "strong growth" assumption motivated by over-parametrized problems and used in [49] to derive an improved convergence rate for AdaGrad is essentially subsumed (at the cumulative level) if "preconditioned cumulative strong growth" (that is $\nu_k = 0$) is assumed in (3.13).

We now start exploiting this result by deriving a more specific bound on the value of Δ_k in (3.14). This requires the following simple technical lemma.

Lemma 3.6 Let $\Gamma \in \mathbb{R}^{d\times d}$ be symmetric positive definite. Then

$$\mathrm{tr}(\log\Gamma) \le 2d\log\big(\mathrm{tr}(\Gamma^{1/2})\big) - d\log(d).$$

Proof. Let $\lambda_1,\ldots,\lambda_d > 0$ be the eigenvalues of Γ. Then the concavity of the logarithm ensures that

$$\mathrm{tr}(\log\Gamma) = \sum_{i=1}^{d}\log(\lambda_i) \le d\log\left(\frac{1}{d}\sum_{i=1}^{d}\lambda_i\right) \le d\log\left(\frac{\left(\sum_{i=1}^{d}\sqrt{\lambda_i}\right)^2}{d}\right) = 2d\log\big(\mathrm{tr}(\Gamma^{1/2})\big) - d\log(d).$$

□

We now derive the desired bound on Δ_k in (3.14).

Lemma 3.7 We have that, for all $k \ge 0$,

$$\Delta_k \le \kappa_d + 2N\log\left(\mathbb{E}\left[\sum_{\ell=1}^{L}\mathrm{tr}(\Gamma_{k,\ell}^{1/2})\right]\right), \tag{3.24}$$

where

$$\kappa_d = -\sum_{\ell=1}^{L} d_\ell\log(d_\ell) - \log(\varsigma)N. \tag{3.25}$$

Proof. Applying Lemma 3.6 to each block gives

$$\sum_{\ell=1}^{L}\mathrm{tr}(\log\Gamma_{k,\ell}) \le 2\sum_{\ell=1}^{L} d_\ell\log\big(\mathrm{tr}(\Gamma_{k,\ell}^{1/2})\big) - \sum_{\ell=1}^{L} d_\ell\log d_\ell. \tag{3.26}$$

Since $\mathrm{tr}(\Gamma_{k,\ell}^{1/2}) \le \sum_{\ell=1}^{L}\mathrm{tr}(\Gamma_{k,\ell}^{1/2})$, we have that $\log\big(\mathrm{tr}(\Gamma_{k,\ell}^{1/2})\big) \le \log\left(\sum_{\ell=1}^{L}\mathrm{tr}(\Gamma_{k,\ell}^{1/2})\right)$, and hence, from (3.26),

$$\sum_{\ell=1}^{L}\mathrm{tr}(\log\Gamma_{k,\ell}) \le 2\sum_{\ell=1}^{L} d_\ell\log\left(\sum_{\ell=1}^{L}\mathrm{tr}(\Gamma_{k,\ell}^{1/2})\right) - \sum_{\ell=1}^{L} d_\ell\log d_\ell.$$

We therefore obtain from (3.14) that

$$\Delta_k \leq -\sum_{\ell=1}^{L} d_\ell \log d_\ell - \sum_{\ell=1}^{L} \mathrm{tr}(\log \Gamma_{-1,\ell}) + 2\sum_{\ell=1}^{L} d_\ell \, \mathbb{E}\left[\log\left(\sum_{\ell=1}^{L} \mathrm{tr}(\Gamma_{k,\ell}^{1/2})\right)\right]. \tag{3.27}$$

But Jensen's inequality for the concave $\log(\cdot)$ function gives that

$$\mathbb{E}\left[\log\left(\sum_{\ell=1}^{L} \mathrm{tr}(\Gamma_{k,\ell}^{1/2})\right)\right] \leq \log\left(\mathbb{E}\left[\sum_{\ell=1}^{L} \mathrm{tr}(\Gamma_{k,\ell}^{1/2})\right]\right).$$

Inserting this inequality in (3.27) and using the definition $\Gamma_{-1,\ell} = \varsigma I_\ell$ gives (3.24)-(3.25). □

We still need a simple technical result before concluding.

Lemma 3.8 Suppose that $1 \leq t \leq c\log(t)$ for $c > 0$. Then $t < 2c\log(2c)$.

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 $x = t/c$ then gives that

$$\log(t) \leq \frac{t}{2c} + \log(2c) - 1 < \frac{t}{2c} + \log(2c),$$

which we may substitute in the inequality $2t \leq 2c\log(t)$, yielding that

$$2t < 2c\left(\frac{t}{2c} + \log(2c)\right) = t + 2c\log(2c),$$

and the result follows by substracting t from both sides. □

We now state a first important result on the expected size of the preconditioners $\Gamma_{k,\ell}$.

Theorem 3.9 Suppose that the ADPREC algorithm is applied to problem (1.1). Suppose that (3.13) and Assumptions 1, 2 and 3 hold. Then, for all $k \geq 0$,

$$\mathbb{E}\left[\sum_{\ell=1}^{L} \mathrm{tr}(\Gamma_{k,\ell}^{1/2})\right] \leq \Theta_k \stackrel{\rm def}{=} \max\left[e^{\max\left[1,\frac{\kappa_d}{2N}\right]}, 3\kappa_{\rm gap}, T_k, Y\right], \tag{3.28}$$

with $\kappa_{\rm gap}$ defined in (3.16),

$$T_k = 12\sqrt{N}\,\nu_k \sqrt{\max\left[1, \log\left(12\sqrt{N}\,\nu_k\right)\right]}, \quad Y = 24N\left(\omega + \frac{L_G\eta}{2}\right)\log\left(24N\left(\omega + \frac{L_G\eta}{2}\right)\right)$$

ν_k being defined in (3.13) and κ_d in (3.25).

Proof. Define

$$t_k = \mathbb{E}\left[\sum_{\ell=1}^{L} \mathrm{tr}(\Gamma_{k,\ell}^{1/2})\right]. \tag{3.29}$$

Substituting (3.24) into (3.15) and using this definition then gives that

$$t_k \leq \kappa_{\rm gap} + \nu_k \sqrt{\kappa_d + 2N \log\left(t_k\right)} + \left(\omega + \frac{L_G \eta}{2}\right)\left(\kappa_d + 2N \log\left(t_k\right)\right). \tag{3.30}$$

Suppose first that

$$\log\left(t_k\right) < \max\left[\frac{1}{2N}, \frac{\kappa_d}{2N}\right].$$

Then

$$t_k \leq e^{\max\left[\frac{1}{2N}, \frac{\kappa_d}{2N}\right]}. \tag{3.31}$$

Alternatively, suppose now that

$$2N \log\left(t_k\right) \geq \max[1, \kappa_d]. \tag{3.32}$$

Then

$$\kappa_d + 2N \log\left(t_k\right) \leq 4N \log\left(t_k\right)$$

and (3.30) can then be rewritten as

$$t_k \leq a + b\, \nu_k \sqrt{\log(t_k)} + c \log(t_k) \tag{3.33}$$

with

$$a = \kappa_{\rm gap}, \quad b = 2\sqrt{N} \quad \text{and} \quad c = 4N\left(\omega + \frac{L_G \eta}{2}\right). \tag{3.34}$$

This formulation is analyzed by distinguishing three cases.
• The first case is when $\max[a, b\nu_k\sqrt{\log(t_k)}, c\log(t_k)] = a$. Then, clearly,

$$t_k \leq 3a. \tag{3.35}$$

• The second case is when

$$\max[a, b\nu_k\sqrt{\log(t_k)}, c\log(t_k)] = b\nu_k\sqrt{\log(t_k)}. \tag{3.36}$$

If $\log(t_k) \leq 1$, then $t_k \leq e$. Alternatively, if $\log(t_k) > 1$ (implying that $t_k > 1$), we deduce that $t_k^2 \leq 9b^2\nu_k^2 \log(t_k)$, that is

$$h_k(t_k) \stackrel{\rm def}{=} \frac{t_k^2}{\tau_k^2} - \log(t_k) \leq 0 \quad \text{with} \quad \tau_k = 3b\nu_k. \tag{3.37}$$

Let $T_k = 2\tau_k\sqrt{\max[1, \log(2\tau_k)]} \geq 2\tau_k$. Then

$$h_k(T_k) = 4\max[1, \log(2\tau_k)] - \log(2\tau_k) - \frac{1}{2}\log(\max[1, \log(2\tau_k)]) > 0. \tag{3.38}$$

Moreover, $h_k'(t) = \dfrac{2t}{\tau_k^2} - \dfrac{1}{t}$, and therefore, for all $t \geq 2\tau_k$, $h_k'(t) \geq \dfrac{4\tau_k}{\tau_k^2} - \dfrac{1}{2\tau_k} = \dfrac{4}{\tau_k} - \dfrac{1}{2\tau_k} > 0$. As a consequence, $h_k(t)$ is increasing for $t \geq 2\tau_k$ and thus $h_k(t) > 0$ for $t \geq T_k$ because of (3.38). We may then deduce from (3.37) that $t_k < T_k$. Thus,

$$t_k \leq \max\left[e, T_k\right]. \tag{3.39}$$

• Finally, the third case is when $\max[a, b\nu_k\sqrt{\log(t_k)}, c\log(t_k)] = c\log(t_k)$. Then $t_k \leq 3c\log(t_k)$ and thus, using Lemma 3.8,

$$t_k \leq \max[1, 6c\log(6c)]. \tag{3.40}$$

Combining (3.29), (3.35), (3.39) and (3.40) and including (3.31) then gives (3.28). □

This result finally gives us all the ingredients to derive convergence and complexity results on the Cartesian product space. The bound (3.28), together with Assumption 4, indeed implies convergence of the criticality measure on each block with a *uniform rate-of-convergence bound*.

Theorem 3.10 Suppose that the ADPREC algorithm is applied to problem (1.1). Suppose that (3.13) and Assumptions 1 to 4 hold. Then

$$\sum_{j=0}^{k} \mathbb{E}\Big[\|\widetilde{G}_j\|_*\Big] \leq \kappa_\circ \sqrt{k+1}\, \Theta_k \tag{3.41}$$

where Θ_k is defined in Theorem 3.9. Moreover, if $\widetilde{G}_k$ is an unbiased estimator of G_k, i.e. $\mathbb{E}_k\Big[\widetilde{G}_k\Big] = G_k$ for all $k \geq 0$, then

$$\min_{j\in\{0,\ldots,k\}} \mathbb{E}[\|G_j\|_*] \leq \frac{1}{k+1}\sum_{j=0}^{k} \mathbb{E}[\|G_j\|_*] \leq \frac{\kappa_\circ\, \Theta_k}{\sqrt{k+1}}. \tag{3.42}$$

Proof. The Cauchy-Schwarz inequality and Assumption 4 imply that

$$\begin{aligned}
\sum_{j=0}^{k} \|\widetilde{G}_j\|_* &\leq \sqrt{k+1}\left(\sum_{j=0}^{k} \|\widetilde{G}_j\|_*^2\right)^{1/2} \\
&= \sqrt{k+1}\left(\sum_{j=0}^{k}\sum_{\ell=1}^{L} \|\widetilde{G}_{j,\ell}\|_{*,\ell}^2\right)^{1/2} \\
&\leq \kappa_\circ \sqrt{k+1}\left(\sum_{j=0}^{k}\sum_{\ell=1}^{L} \mathrm{tr}\big(\mathcal{U}_{j,\ell}(\widetilde{G}_{j,\ell})\big)\right)^{1/2}.
\end{aligned}$$

Now, (2.4) ensures that

$$\sum_{j=0}^{k} \mathcal{U}_{j,\ell}(\widetilde{G}_{j,\ell}) = \Gamma_{k,\ell} - \Gamma_{-1,\ell},$$

so that

$$\sum_{j=0}^{k} \|\widetilde{G}_j\|_* \leq \kappa_\circ \sqrt{k+1}\left(\sum_{\ell=1}^{L} \mathrm{tr}(\Gamma_{k,\ell})\right)^{1/2} \leq \kappa_\circ \sqrt{k+1}\sum_{\ell=1}^{L} \mathrm{tr}(\Gamma_{k,\ell}^{1/2}).$$

Taking the total expectation on both sides of this inequality and using Theorem 3.9 then yields that

$$\sum_{j=0}^{k} \mathbb{E}\Big[\|\widetilde{G}_j\|_*\Big] \leq \kappa_\circ \sqrt{k+1}\, \mathbb{E}\left[\sum_{\ell=1}^{L} \mathrm{tr}(\Gamma_{k,\ell}^{1/2})\right] \leq \kappa_\circ \sqrt{k+1}\, \Theta_k,$$

proving (3.41). Dividing by $k+1$ and using $\mathbb{E}_k\Big[\widetilde{G}_k\Big] = G_k$ together with Jensen's inequality gives (3.42). □

We now investigate what can be said if one makes a more specific assumption on conditional variance at each iteration.

Corollary 3.11 Suppose that the ADPREC algorithm is applied to problem (1.1). Suppose that Assumptions 1 to 4 hold. Suppose also that

$$\mathbb{E}_k\Big[\widetilde{G}_k\Big] = G_k \quad \text{and} \quad \mathbb{E}_k\Big[\|\widetilde{G}_{k,\ell} - G_{k,\ell}\|_{*,\ell}^2\Big] \leq \frac{\sigma_\ell^2}{(k+1)^\alpha} + \omega^2 \mathbb{E}_k\big[\|Z_{k,\ell}\|_{*,\ell}^2\big] \tag{3.43}$$

for some $\alpha, \omega > 0$ and all $k \geq 0$ and $\ell \in \{1, \ldots, L\}$. Then

$$\frac{1}{k+1}\sum_{j=0}^{k} \mathbb{E}[\|G_j\|_*] = \begin{cases} \mathcal{O}\left(\dfrac{\sqrt{\log(k+1)}}{(k+1)^{\frac{\alpha}{2}}}\right) & \text{if} \quad \alpha < 1, \\ \mathcal{O}\left(\dfrac{\sqrt{\log(k+1)\log(\log(k+1))}}{\sqrt{k+1}}\right) & \text{if} \quad \alpha = 1, \\ \mathcal{O}\left(\dfrac{1}{\sqrt{k+1}}\right) & \text{if} \quad \alpha > 1. \end{cases} \tag{3.44}$$

Proof. Observe first that the term in the expression of Θ_k in (3.28) with potential fastest growth when k tends to infinity is T_k, yielding (in order)

$$\Theta_k = \mathcal{O}\left(\nu_k \sqrt{\log(\nu_k)}\right). \tag{3.45}$$

But (3.13) and (3.43) imply that

$$\nu_k^2 = \sigma_{\rm tot}^2 \sum_{j=0}^{k} \frac{1}{(j+1)^\alpha},$$

where $\sigma_{\rm tot}^2 = \sum_{\ell=1}^{L} \sigma_\ell^2$. Suppose first that $\alpha < 1$. We then have that $\sum_{j=0}^{k} \frac{1}{(j+1)^\alpha} \leq (k+1)^{1-\alpha}/(1-\alpha)$ so that $\nu_k = \frac{\sigma_{\rm tot}}{\sqrt{1-\alpha}}\,(k+1)^{\frac{1}{2}(1-\alpha)}$ in this case. If $\alpha = 1$, then $\sum_{j=0}^{k} \frac{1}{(j+1)^\alpha} = \sum_{j=0}^{k} \frac{1}{j+1} \leq 1 + \log(k+1)$ and thus $\nu_k \leq \sigma_{\rm tot}\sqrt{1+\log(k+1)}$. Finally, if $\alpha > 1$, we have that $\sum_{j=0}^{k} \frac{1}{(j+1)^\alpha} \leq \zeta(\alpha) < +\infty$, where $\zeta(\cdot)$ is the Riemann zeta function. Combining the three cases, we obtain that

$$\nu_k \leq \begin{cases} \dfrac{\sigma_{\rm tot}}{\sqrt{1-\alpha}}(k+1)^{\frac{1}{2}(1-\alpha)} & \text{if} \quad \alpha < 1, \\ \sigma_{\rm tot}\sqrt{1+\log(k+1)} & \text{if} \quad \alpha = 1, \\ \sigma_{\rm tot}\zeta(\alpha) & \text{if} \quad \alpha > 1, \end{cases} \tag{3.46}$$

The rates (3.44) then follow from these bounds and (3.42). □

Note that the continuity of the bound expressed in Theorem 3.10 as a function of ν_k is lost in the statement of Corollary 3.11, because, as is clear from its proof, the constants hidden in the $\mathcal{O}(\cdot)$ notation depend on α. In particular, the formulation using $\mathcal{O}(\cdot)$ does not allow taking the limit for α tending to one.

Observe that we could replace the second part of (3.43) by

$$\mathbb{E}_k\Big[\|\widetilde{G}_{k,\ell} - G_{k,\ell}\|_{*,\ell}^2\Big] \leq \frac{\sigma_\ell^2}{(k+1)^\alpha} + \omega^2 \|Z_{k-1,\ell}\|_{*,\ell}^2 \tag{3.47}$$

with the convention that $Z_{-1,\ell} = 0$, since

$$\sum_{j=0}^{k-1}\sum_{\ell=1}^{L} \mathbb{E}_j\big[\|Z_{j,\ell}\|_{*,\ell}^2\big] \leq \sum_{j=0}^{k}\sum_{\ell=1}^{L} \mathbb{E}_j\big[\|Z_{j,\ell}\|_{*,\ell}^2\big].$$

The advantage of (3.47) is that the right-hand-side of this new condition is now $\mathcal{F}_k$-measurable at iteration k while this is not the case for the second part of (3.43) because the $Z_{k,\ell}$ are not $\mathcal{F}_k$-measurable.

Although the rates of convergence obtained in Corollary 3.11 under the bias and variance conditions (3.43) are reasonable, they do not quite match, for high-variance regimes, those obtained for (momentum-less) AdaGrad and AdaNorm [49], which do not depend on the finiteness of the "inaccuracy budget" ν_k^2 and thus allow for bounded conditional variance at every iteration (condition (3.13) may require increasing batch-size) . However they recover (in order) the best[3] $\mathcal{O}(\sqrt{\log(k+1)\log(\log(k+1))/(k+1)})$ rate obtained in [49] in the preconditioned cumulative strong growth context[4]. Importantly, they do so for a large class of algorithms. While it would be desirable to derive general convergence results under the standard "bounded variance" assumption (that is $\alpha = \omega = 0$ in (3.43)), this turns out to be impossible for a unified theory covering Muon, as shown by the following simple example. Consider using Muon (see Section(2.4)) to minimize

$$f(X) = \frac{1}{4}\big([X]_{1,1} - [X]_{2,2}\big)^2 \quad \text{starting from} \quad X_0 = \mathrm{diag}(1,-1) \in \mathbb{R}^{2\times 2}.$$

We have that $G(X) = \mathrm{diag}(r(X), -r(X))$ with $r(X) = \frac{1}{2}([X]_{1,1} - [X]_{2,2})$, so that $r(X_0) = 2$ and $\|G(X_0)\|_* = 1$ where $\|\cdot\|_*$ is the nuclear norm. At iteration k, let $\widetilde{G}_k = \mathrm{diag}(3\,r(X_k), r(X_k))$ or $\widetilde{G}(X) = \mathrm{diag}(-r(X_k), 3\,r(X_k))$ with equal probabilities. Then $\mathbb{E}_k\Big[\widetilde{G}_k\Big] = G_k$, $\mathbb{E}_k\Big[\|\widetilde{G}_k - G_k\|_*^2\Big] = 16\,r(X_k)^2$, and one verifies that $\mathcal{N}_1(\widetilde{G}_k)$ given by (2.10) is then either the identity or its opposite with equal probabilities. Thus every Muon step is a multiple of the identity matrix, which leaves $r(X)$, and therefore $G(X)$, unchanged. As a consequence, $\|G(X_k)\|_* = 1$ for all $k \geq 0$, despite $\mathbb{E}_k\Big[\|\widetilde{G}_k - G_k\|_*^2\Big] = 64$ remaining bounded for all k. We also note that, should the restrictive assumption of a bounded gradient oracle be made for AdaGrad, only the first part of (3.43) (unbiased oracle) is necessary to deduce that the average of $\mathbb{E}\big[\|G_k\|_E^2\big]$ decreases in $\mathcal{O}(1/(k+1)^{1/4})$ [21].

3.3 Adding momentum (twice)

3.3.1 Using a momentum-based preconditioner

Because of its widespread use, we now consider properties of variants of ADPREC augmented with momentum, and start with the ADPREC.MM given on the next page.

This is ADPREC where (2.5) and (2.4) are replaced by (3.48)–(3.50). This modification of the algorithm in effect adds a specific type of momentum, although this is not the only possible one. Another variant where $\mathcal{U}_{k,\ell}(\widetilde{G}_{k,\ell})$ is used instead of $\mathcal{U}_{k,\ell}(M_{k,\ell})$ will be considered in Section 3.3.2.

We now show that the theory of the previous section can be extended to cover ADPREC.MM. We start by establishing a bound on the norm of the difference between the block-wise momentum $M_{k,\ell}$ and the approximate gradient $\widetilde{G}_{k,\ell}$.

[3] It is not known whether a better rate is achievable for AdaGrad and Adanorm.

[4] That is not only in the case where $\sigma_\ell = 0$ for each ℓ, but, more generally, in the case where the "inaccuracy budget" ν_k^2 is finite.

Algorithm 3.1: ADPREC.MM

Given: a starting point X_0, operators $\{\mathcal{U}_{k,\ell}\}_{k\geq 0,\ell\in\{1,\ldots,L\}}$ and $\{\mathcal{N}_\ell\}_{\ell\in\{1,\ldots,L\}}$, and constants $\eta,\varsigma>0$ and $\mu_{\max}\in(0,1)$. Set $\Gamma_{-1,\ell}=\varsigma I_\ell$ for $\ell\in\{1,\ldots,L\}$.

For $k=0,1,\ldots$

1. Draw ξ_k and compute $\widetilde{G}_k=\widetilde{G}(X_k,\xi_k)=(\widetilde{G}_{k,1},\ldots,\widetilde{G}_{k,L})$, [gradient approx.]
2. On each block $\ell\in\{1,\ldots,L\}$, compute

$$M_{k,\ell} = \begin{cases} \widetilde{G}_{0,\ell} & \text{if } k=0,\\ \mu_k M_{k-1,\ell}+(1-\mu_k)\widetilde{G}_{k,\ell} & \text{if } k>0, \end{cases} \quad \text{[momentum M]} \tag{3.48}$$

$$\Gamma_{k,\ell} = \Gamma_{k-1,\ell}+\mathcal{U}_{k,\ell}(M_{k,\ell}), \quad \text{[preconditioner update]} \tag{3.49}$$

$$Z_{k,\ell} = \Gamma_{k,\ell}^{-1/2}M_{k,\ell}, \quad \text{[preconditioned gradient]} \tag{3.50}$$

$$X_{k+1,\ell} = X_{k,\ell}-\eta\,\|Z_{k,\ell}\|_{*,\ell}\mathcal{N}_\ell(Z_{k,\ell}). \quad \text{[scaled normalized step]} \tag{3.51}$$

Lemma 3.12 Suppose that Assumption 2 holds. Define

$$E_{k,\ell}=M_{k,\ell}-\widetilde{G}_{k,\ell} \quad \text{for} \quad k\geq 0 \ \text{ and } \ \ell\in\{1,\ldots,L\}. \tag{3.52}$$

Then

$$\sum_{j=0}^{k}\mathbb{E}\big[\|E_j\|_*^2\big]\leq\frac{6}{(1-\mu_{\max})^2}\sum_{j=0}^{k}\bar{\mu}_j^2\mathbb{E}\Big[\|\widetilde{G}_j-G_j\|_*^2\Big]+\frac{3L_G^2\eta^2}{(1-\mu_{\max})^2}\sum_{j=0}^{k}\bar{\mu}_j^2\mathbb{E}\big[\|Z_j\|_*^2\big], \tag{3.53}$$

where

$$\bar{\mu}_j=\max[\mu_j,\mu_{j+1}]. \tag{3.54}$$

Proof. If $k=0$, (3.53) trivially holds. Suppose that $k\geq 1$ and note that (3.52) gives that

$$\begin{aligned} E_{k,\ell} &= \mu_k M_{k-1,\ell}+(1-\mu_k)\widetilde{G}_{k,\ell}-\widetilde{G}_{k,\ell}\\ &= \mu_k(E_{k-1,\ell}+\widetilde{G}_{k-1,\ell})+(1-\mu_k)\widetilde{G}_{k,\ell}-\widetilde{G}_{k,\ell}\\ &= \mu_k E_{k-1,\ell}+\mu_k(\widetilde{G}_{k-1,\ell}-\widetilde{G}_{k,\ell}). \end{aligned}$$

Using the facts that $\mu_k\leq\mu_{\max}$, that $(a+b)^2\leq(1+\tau)a^2+(1+1/\tau)b^2$ for any $\tau>0$ and that $(a+b+c)^2\leq 3a^2+3b^2+3c^2$, we obtain that

$$\begin{aligned} \|E_{k,\ell}\|_{*,\ell}^2\leq&(1+\tau)\mu_{\max}^2\|E_{k-1,\ell}\|_{*,\ell}^2\\ &+3\mu_k^2\left(1+\frac{1}{\tau}\right)\left(\|G_{k,\ell}-G_{k-1,\ell}\|_{*,\ell}^2+\|\widetilde{G}_{k,\ell}-G_{k,\ell}\|_{*,\ell}^2+\|\widetilde{G}_{k-1,\ell}-G_{k-1,\ell}\|_{*,\ell}^2\right) \end{aligned}$$

Summing over $\ell\in\{1,\ldots,L\}$ and using (2.2), we obtain that

$$\begin{aligned} \|E_k\|_*^2\leq&(1+\tau)\mu_{\max}^2\|E_{k-1}\|_*^2\\ &+3\mu_k^2\left(1+\frac{1}{\tau}\right)\left(\|G_k-G_{k-1}\|_*^2+\sum_{\ell=1}^{L}\|\widetilde{G}_{k,\ell}-G_{k,\ell}\|_{*,\ell}^2+\sum_{\ell=1}^{L}\|\widetilde{G}_{k-1,\ell}-G_{k-1,\ell}\|_{*,\ell}^2\right) \end{aligned}$$

Observe now that Assumption 2 with (3.51) and (2.3) imply that

$$\|G_k - G_{k-1}\|_*^2 \leq L_G^2 \sum_{\ell=1}^{L} \|X_{k,\ell} - X_{k-1,\ell}\|^2 = L_G^2 \eta^2 \sum_{\ell=1}^{L} \|Z_{k-1,\ell}\|_{*,\ell}^2 \|\mathcal{N}_\ell(Z_{k-1,\ell})\|_\ell^2 = L_G^2 \eta^2 \|Z_{k-1}\|_*^2, \tag{3.55}$$

Now let $\tau = (1-\mu_{\max})/\mu_{\max}$. Then $(1+\tau)\mu_{\max}^2 = \mu_{\max} < 1$ and $(1+1/\tau)\mu_k^2 = \mu_k^2/(1-\mu_{\max})$. If $\vartheta_k^2 = \sum_{\ell=1}^{L} \mathbb{E}\Big[\|\widetilde{G}_{k,\ell} - G_{k,\ell}\|_{*,\ell}^2\Big] = \mathbb{E}\Big[\|\widetilde{G}_k - G_k\|_*^2\Big]$, we have, after taking total expectation and using (3.55) that

$$\mathbb{E}\big[\|E_k\|_*^2\big] \leq \mu_{\max} \mathbb{E}\big[\|E_{k-1}\|_*^2\big] + \frac{3\mu_k^2}{1-\mu_{\max}} \left(L_G^2 \eta^2 \mathbb{E}\big[\|Z_{k-1}\|_*^2\big] + \vartheta_k^2 + \vartheta_{k-1}^2\right)$$

By the definition of $\bar{\mu}_j$,

$$\sum_{j=1}^{k} \mu_j^2 \mathbb{E}\big[\|Z_{j-1}\|_*^2\big] = \sum_{j=0}^{k-1} \mu_{j+1}^2 \mathbb{E}\big[\|Z_j\|_*^2\big] \leq \sum_{j=0}^{k} \bar{\mu}_j^2 \mathbb{E}\big[\|Z_j\|_*^2\big].$$

Moreover,

$$\sum_{j=1}^{k} \mu_j^2 \left(\vartheta_j^2 + \vartheta_{j-1}^2\right) = \sum_{j=1}^{k} \mu_j^2 \vartheta_j^2 + \sum_{j=0}^{k-1} \mu_{j+1}^2 \vartheta_j^2 \leq 2 \sum_{j=0}^{k} \bar{\mu}_j^2 \vartheta_j^2.$$

Combining these inequalities therefore gives

$$\sum_{j=0}^{k} \mathbb{E}\big[\|E_j\|_*^2\big] \leq \frac{6}{(1-\mu_{\max})^2} \sum_{j=0}^{k} \bar{\mu}_j^2 \mathbb{E}\big[\|\widetilde{G}_j - G_j\|_*^2\big] + \frac{3L_G^2\eta^2}{(1-\mu_{\max})^2} \sum_{j=0}^{k} \bar{\mu}_j^2 \mathbb{E}\big[\|Z_j\|_*^2\big],$$

which proves (3.53). □

We now wish to apply the theory above by considering that the momentum $M_{k,\ell}$ plays the role of the approximate gradient (in Step 2 of ADPREC) in this theory. In particular, this change of perspectives implies that Assumption 3 now becomes

Assumption 5 (Modified structural identities) *For all $k \geq 0$ and all $\ell \in \{1, \ldots, L\}$, we have that*

$$\|Z_{k,\ell}\|_{*,\ell} \, \langle M_{k,\ell}, \mathcal{N}_\ell(Z_{k,\ell})\rangle_F = \mathrm{tr}\big(\Gamma_{k,\ell}^{-1/2} \mathcal{U}_{k,\ell}(M_{k,\ell})\big), \tag{3.56}$$

and

$$\|Z_{k,\ell}\|_{*,\ell}^2 = \mathrm{tr}\big(\Gamma_{k,\ell}^{-1} \mathcal{U}_{k,\ell}(M_{k,\ell})\big). \tag{3.57}$$

Fortunately, using our theory is possible because, as we show below, (3.53) implies that condition (3.13) of Theorem 3.5 holds with suitable constants. This allows us to derive the analog of Theorem 3.10 for the "momentum" variant of ADPREC.

Theorem 3.13 Suppose that the ADPREC algorithm modified by (3.48)–(3.50) is applied to problem (1.1) and that Assumptions 1, 2, 4 and 5 hold. Then

$$\frac{1}{k+1} \sum_{j=0}^{k} \mathbb{E}[\|G_j\|_*] \leq \frac{1}{\sqrt{k+1}} \Big((\kappa_\circ + 1)\Theta_k + \omega\sqrt{2N\log(\Theta_k)} + \omega\sqrt{\max[\kappa_d, 1]}\Big) \tag{3.58}$$

where Θ_k is defined in (3.28)–(3.25) using

$$\nu_k^2 = \left(\frac{12\,\mu_{\max}^2}{(1-\mu_{\max})^2} + 2\right) \sum_{j=0}^{k} \mathbb{E}\Big[\|\widetilde{G}_j - G_j\|_*^2\Big] \quad \text{and} \quad \omega^2 = \frac{6\,\mu_{\max}^2 L_G^2 \eta^2}{(1-\mu_{\max})^2}. \tag{3.59}$$

Proof. Observe first that the identity $(a+b)^2 \leq 2a^2+2b^2$, (3.53) and the bound $\bar{\mu}_k \leq \mu_{\max}$ give that

$$\begin{aligned}\sum_{j=0}^{k} \mathbb{E}\big[\|M_j - G_j\|_*^2\big] &\leq 2\sum_{j=0}^{k} \mathbb{E}\Big[\|\widetilde{G}_j - G_j\|_*^2\Big] + 2\sum_{j=0}^{k} \mathbb{E}\big[\|E_j\|_*^2\big] \\ &\leq \left(\frac{12\,\mu_{\max}^2}{(1-\mu_{\max})^2} + 2\right) \sum_{j=0}^{k} \mathbb{E}\Big[\|\widetilde{G}_j - G_j\|_*^2\Big] + \frac{6\,\mu_{\max}^2 L_G^2 \eta^2}{(1-\mu_{\max})^2} \sum_{j=0}^{k} \mathbb{E}\big[\|Z_j\|_*^2\big]\end{aligned} \tag{3.60}$$

which is (3.13) with (3.59). We may therefore apply Theorem 3.10 (under Assumption 5) with these values and deduce that

$$\sum_{j=0}^{k} \mathbb{E}[\|M_j\|_*] \leq \sqrt{k+1}\,\kappa_\circ\,\Theta_k. \tag{3.61}$$

Hence, using the triangle, Cauchy-Schwarz and Jensen inequalities, we obtain that

$$\begin{aligned}\sum_{j=0}^{k} \mathbb{E}[\|G_j\|_*] &\leq \sum_{j=0}^{k} \mathbb{E}[\|M_j\|_*] + \sum_{j=0}^{k} \mathbb{E}[\|M_j - G_j\|_*] \\ &\leq \sum_{j=0}^{k} \mathbb{E}[\|M_j\|_*] + \sqrt{k+1}\sqrt{\sum_{j=0}^{k} \mathbb{E}[\|M_j - G_j\|_*]^2} \\ &\leq \sqrt{k+1}\,\kappa_\circ\,\Theta_k + \sqrt{k+1}\sqrt{\sum_{j=0}^{k} \mathbb{E}[\|M_j - G_j\|_*^2]} \\ &\leq \sqrt{k+1}\,\kappa_\circ\,\Theta_k + \sqrt{k+1}\sqrt{\nu_k^2 + \omega^2 \sum_{j=0}^{k}\sum_{\ell=1}^{L} \mathbb{E}\Big[\|Z_{j,\ell}\|_{*,\ell}^2\Big]}\end{aligned} \tag{3.62}$$

where we have inserted (3.59) in (3.60) to obtain the last equality. But, successively using (3.20), (3.24) and (3.28), we deduce that

$$\sum_{j=0}^{k}\sum_{\ell=1}^{L} \mathbb{E}\big[\|Z_{j,\ell}\|_{*,\ell}^2\big] \leq \Delta_k \leq \kappa_d + 2N\log\left(\mathbb{E}\left[\sum_{\ell=1}^{L} \mathrm{tr}(\Gamma_{k,\ell}^{1/2})\right]\right) \leq \kappa_d + 2N\log\left(\Theta_k\right),$$

Substituting this inequality into (3.62) gives

$$\begin{aligned}\sum_{j=0}^{k} \mathbb{E}[\|G_j\|_*] &\leq \sqrt{k+1}\,\kappa_\circ\Theta_k + \sqrt{k+1}\sqrt{\nu_k^2 + \omega^2\big(\kappa_d + 2N\log(\Theta_k)\big)} \\ &\leq \sqrt{k+1}\left[\kappa_\circ\Theta_k + \nu_k + \omega\sqrt{\kappa_d + 2N\log(\Theta_k)}\right].\end{aligned}$$

Since $\nu_k \leq \Theta_k$, we obtain

$$\sum_{j=0}^{k} \mathbb{E}[\|G_j\|_*] \leq \sqrt{k+1}\left[(\kappa_\circ+1)\Theta_k + \omega\sqrt{\kappa_d + 2N\log(\Theta_k)}\right].$$

Dividing by $k+1$ and using $\sqrt{a+b} \leq \sqrt{a} + \sqrt{b}$ therefore yields (3.58). □

The difference between (3.42) and (3.58) is only a constant and a logarithmic term in factor of $(k+1)^{-1/2}$. As a consequence, the global rate of convergence of the momentum variant of ADPREC

is essentially identical to that of the version without momentum, and Corollary 3.11 still applies to the momentum variant.

We finally note that our momentum definition and convergence analysis do not make the assumption that the stepsize parameter η is sufficiently small, in contrast with previous proofs of convergence where results assume that η is small enough, in particular smaller than a multiple of the (usually unknown) Lipschitz constant L_G [23, 26, 53].

3.3.2 Using a gradient based preconditioner

The momentum approach we have described above uses (3.48)–(3.50). Another version can be considered, using (3.48) and (3.50) but keeping the "pure gradient" technique (2.4) to accumulate the preconditioner, as shown on this page.

Algorithm 3.2: ADPREC.MG

Given: a starting point X_0 operators $\{\mathcal{U}_{k,\ell}\}_{k\geq 0,\ell\in\{1,\ldots,L\}}$ and $\{\mathcal{N}_\ell\}_{\ell\in\{1,\ldots,L\}}$, and constants $\eta,\varsigma>0$ and $\mu_{\max}\in(0,1)$. Set $\Gamma_{-1,\ell}=\varsigma I_\ell$ for $\ell\in\{1,\ldots,L\}$.

For $k=0,1,\ldots$

1. Draw ξ_k and compute $\widetilde{G}_k=\widetilde{G}(X_k,\xi_k)=(\widetilde{G}_{k,1},\ldots,\widetilde{G}_{k,L})$, [gradient approx.]
2. On each block $\ell\in\{1,\ldots,L\}$, compute

$$\Gamma_{k,\ell} = \Gamma_{k-1,\ell}+\mathcal{U}_{k,\ell}(\widetilde{G}_{k,\ell}), \quad \text{[preconditioner update]} \quad (3.63)$$

$$M_{k,\ell} = \begin{cases} \widetilde{G}_{0,\ell} & \text{if } k=0,\\ \mu_k M_{k-1,\ell}+(1-\mu_k)\widetilde{G}_{k,\ell} & \text{if } k>0, \end{cases} \quad \text{[momentum G]} \quad (3.64)$$

$$Z_{k,\ell} = \Gamma_{k,\ell}^{-1/2} M_{k,\ell}, \quad \text{[preconditioned gradient]} \quad (3.65)$$

$$X_{k+1,\ell} = X_{k,\ell}-\eta\,\|Z_{k,\ell}\|_{*,\ell}\,\mathcal{N}_\ell(Z_{k,\ell}). \quad \text{[scaled normalized step]} \quad (3.66)$$

As it turns out, it is also possible to derive a unified convergence theory for this choice, albeit this requires stronger assumptions and leads to a possibly worse rate of convergence. This alternative theory hinges on the fact that $\mathcal{U}_{k,\ell}(\widetilde{G}_{k,\ell})$ can be viewed as a perturbation of $\mathcal{U}_{k,\ell}(M_{k,\ell})$, and therefore that the structural relations of Assumption 5 are themselves perturbed. Fortunately, it remains possible to bound these perturbations by a mix of variance-related and second-order terms, the first of which potentially affecting the resulting global convergence rate. The final outcome is given by the following theorem and corollary (whose detailed proofs can be found in the appendix).

Theorem 3.14 Suppose that the ADPREC algorithm with (2.5) replaced by (3.64) and (3.65) is applied to problem (1.1). Suppose that Assumptions 1, 2, 4 and 5 hold. If $\mu > 0$, suppose also that there exists constants $\kappa_\square, \kappa_\diamond > 0$ such that

$$\mathcal{U}_{k,\ell}(U+V) \preceq \kappa_\square \mathcal{U}_{k,\ell}(U) + \kappa_\square \mathcal{U}_{k,\ell}(V) \quad \text{and} \quad \mathrm{tr}\Big(\mathcal{U}_{k,\ell}(U)\Big) \leq \kappa_\diamond \|U\|^2_{*,\ell} \tag{3.67}$$

for all $\ell \in \{1, \ldots, L\}$, $k \geq 0$ and all $U, V \in \mathbb{R}^{n_\ell \times m_\ell}$, and that

$$\text{either} \quad \eta \leq \frac{1-\mu_{\max}}{\mu_{\max} L_G} \sqrt{\frac{\varsigma}{6\kappa_\square \kappa_\diamond}} \quad \text{or} \quad \sum_{j=0}^{\infty} \bar{\mu}_j^2 \|Z_j\|_*^2 \leq \kappa_{\mu Z} \tag{3.68}$$

for some $\kappa_{\mu Z} \geq 0$ and $\bar{\mu}_j$ defined by (3.54). Then

$$\sum_{j=0}^{k} \mathbb{E}\Big[\|\widetilde{G}_j\|_*\Big] \leq \sqrt{k+1}\, \kappa_\circ\, \Theta_k \tag{3.69}$$

where

$$\Theta_k = \max\Big[\, e^{\max\left[1, \frac{\kappa_d}{2N}\right]},\, 3(\kappa_{\rm gap} + \varepsilon(\nu_k, \theta_k)), T_k,\, Y\,\Big], \tag{3.70}$$

with ν_k defined in (3.13),

$$\varepsilon(\nu, \theta) = \kappa_\square \left(\sqrt{\kappa_{2,z}}\, \nu + \sqrt{\kappa_{2,\theta}}\, \nu\theta + \kappa_{\theta\theta}\theta^2\right)$$

$$\theta_k^2 = \sum_{j=0}^{k} \sum_{\ell=1}^{L} \bar{\mu}_j^2 \mathbb{E}\Big[\|\widetilde{G}_{j,\ell} - G_{j,\ell}\|^2_{*,\ell}\Big], \tag{3.71}$$

$$T_k = 6(2\kappa_\square)^{3/2} \sqrt{N}\nu_k \sqrt{\max\Big[1, \log\Big(6(2\kappa_\square)^{3/2}\sqrt{N}\ \nu_k\Big)\Big]},$$

$$Y = 24N\kappa_\Delta \kappa_\square^2 \log\left(24N\kappa_\Delta\kappa_\square^2\right),$$

$\kappa_\circ$ being defined in Assumption 4, κ_d in (3.25), $\kappa_{2,z}$ and $\kappa_{2,\theta}$ in (A.11) and $\kappa_{\rm gap}$, $\kappa_{\theta\theta}$ and κ_Δ in (A.13)–(A.15). Moreover, if $\widetilde{G}_k$ is an unbiased estimator of G_k, i.e. $\mathbb{E}_k\Big[\widetilde{G}_k\Big] = G_k$ for all $k \geq 0$, then

$$\min_{j\in\{0,\ldots,k\}} \mathbb{E}[\|G_j\|_*] \leq \frac{1}{k+1} \sum_{j=0}^{k} \mathbb{E}[\|G_j\|_*] \leq \frac{\kappa_\circ\, \Theta_k}{\sqrt{k+1}}. \tag{3.72}$$

As it turns out, assuming (3.67) is not restrictive in the applications considered in this paper, as we briefly discuss below in the appendix. Supposing (3.68) is definitely less desirable because the value of the Lipschitz constant L_G or that of $\kappa_{\mu Z}$ (if it exists) is usually unknown. It is however common practice in the literature [23, 26, 53]. Note the term $\varepsilon(\nu_k, \theta_k)$ in the middle term of (3.70), which is the crucial difference between (3.70) and (3.28). Because $\theta_k < \nu_k$ (since $\mu_k \leq \mu_{\max} < 1$), the term which achieves the fastest growth when k tends to infinity in (3.69) is the largest of $\nu_k\theta_k$ (the second term of $\varepsilon(\nu_k, \theta_k)$) and $\nu_k\sqrt{\log(\nu_k)}$ (the T_k term), that is $\nu_k \max[\theta_k, \sqrt{\log(\nu_k)}]$. Thus the rate of convergence of ADPREC.MG is never better than that of ADPREC and is only as good when $\theta_k = \mathcal{O}\left(\sqrt{\log(\nu_k)}\right)$. This induces the modified convergence rate expressed by the following corollary.

Corollary 3.15 Suppose that the ADPREC algorithm with (2.5) replaced by (3.64) and (3.65) is applied to problem (1.1). Suppose that Assumptions 1, 2, 4 and 5 hold. If $\mu > 0$, suppose also that (3.67) and (3.68) hold, that

$$\mathbb{E}_k\Big[\widetilde{G}_k\Big] = G_k \quad \text{and} \quad \mathbb{E}_k\Big[\|\widetilde{G}_{k,\ell} - G_{k,\ell}\|_*^2\Big] \leq \frac{\sigma_\ell^2}{(k+1)^\alpha} \tag{3.73}$$

for some $\alpha > 0$ and all $k \geq 0$ and $\ell \in \{1, \ldots, L\}$, and that

$$\mu_k = \frac{\mu_{\max}}{(k+1)^\beta}$$

for some $\mu_{\max} < 1$ and some $\beta \geq 0$.

$$\frac{1}{k+1}\sum_{j=0}^{k}\mathbb{E}[\|G_j\|_*] = \begin{cases} \mathcal{O}\left(\dfrac{1}{(k+1)^{\alpha+\beta-\frac{1}{2}}}\right) & \text{if} \quad \alpha+2\beta<1, \\ \mathcal{O}\left(\dfrac{\sqrt{\log(k+1)}}{(k+1)^{\frac{\alpha}{2}}}\right) & \text{if} \quad \alpha<1 \ \text{ and } \ \alpha+2\beta\geq 1, \\ \mathcal{O}\left(\dfrac{\log(k+1)}{\sqrt{k+1}}\right) & \text{if} \quad \alpha=1 \ \text{ and } \ \alpha+2\beta=1, \\ \mathcal{O}\left(\dfrac{\sqrt{\log(k+1)\log(\log(k+1))}}{\sqrt{k+1}}\right) & \text{if} \quad \alpha=1 \ \text{ and } \ \alpha+2\beta>1, \\ \mathcal{O}\left(\dfrac{1}{\sqrt{k+1}}\right) & \text{if} \quad \alpha>1. \end{cases} \tag{3.74}$$

Note that, when $\alpha+2\beta < 1$, the right-hand side of (3.74) tends to zero only if $\alpha+\beta > \frac{1}{2}$. One should however be careful not to identify better theoretical convergence properties with better practical performance on specific classes of problems. Indeed, limited numerical experiments suggest that the second momentum variant might outperform the first.

4 Application to existing algorithms

We now consider the four popular first-order algorithms introduced in Sections 2.1 to 2.4 and show that their convergence behaviour is covered by the above results. This is achieved by verifying that the conditions (3.1), (3.2) and (3.3) hold in each case for all blocks concerned. In all cases considered here, the preconditioner update $\mathcal{U}_{k,\ell}$ is independent of k. For simplicity of exposition, we focus on the situation where there is no momentum ($\mu_k = 0$). We again use lower case symbols for vectors.

4.1 AdaNorm [45, 15, 51]

With the choices specified by (2.7), we have that

$$\operatorname{tr}\left(\Gamma_{k,\ell}^{-1}\mathcal{U}_{k,\ell}(\widetilde{G}_{k,\ell})\right) = \operatorname{tr}\left(\frac{\|\widetilde{G}_{k,\ell}\|_\ell^2}{n_\ell}\Gamma_{k,\ell}^{-1}\right).$$

Because of the definition of $\Gamma_{k,\ell}$, this becomes

$$\operatorname{tr}\left(\Gamma_{k,\ell}^{-1}\mathcal{U}_{k,\ell}(\widetilde{G}_{k,\ell})\right) = \frac{\|\widetilde{G}_{k,\ell}\|_\ell^2}{\gamma_{k,\ell}} = \|Z_{k,\ell}\|_\ell^2,$$

which is (3.2). Similarly,

$$\mathrm{tr}(\Gamma_{k,\ell}^{-1/2}\mathcal{U}_{k,\ell}(\widetilde{G}_{k,\ell})) = \frac{\|\widetilde{G}_{k,\ell}\|_\ell^2}{\sqrt{\gamma_{k,\ell}}} = \|Z_{k,\ell}\|_\ell \left\langle \widetilde{G}_{k,\ell}, \mathcal{N}_\ell(Z_{k,\ell})\right\rangle,$$

proving (3.1). Moreover, Assumption 4 also holds on block ℓ. Indeed, for any $\widetilde{G}_\ell$, the second and last parts of (2.7) give

$$\mathrm{tr}(\mathcal{U}_{k,\ell}(\widetilde{G}_\ell)) = \|\widetilde{G}_\ell\|_\ell^2 = \|\widetilde{G}_\ell\|_{*,\ell}^2,$$

and thus (3.3) holds with $\kappa_\circ = 1$.

Thus, whenever a block ℓ is endowed with this isotropic scalar geometry, Theorem 3.10 and Corollary 3.11 apply to that block without modification. The standard "one block" AdaNorm is obtained when there is only one block ($L = 1, n_1 = d_1 = N$).

4.2 Full AdaGrad [15]

With the choices (2.8), we have that

$$\mathrm{tr}(\Gamma_{k,\ell}^{-1}\mathcal{U}_{k,\ell}(\widetilde{G}_{k,\ell})) = \mathrm{tr}(\Gamma_{k,\ell}^{-1/2}\widetilde{G}_{k,\ell}\widetilde{G}_{k,\ell}^T\Gamma_{k,\ell}^{-1/2}) = \|Z_{k,\ell}\|_E^2,$$

which is (3.2). Similarly,

$$\mathrm{tr}(\Gamma_{k,\ell}^{-1/2}\mathcal{U}_{k,\ell}(\widetilde{G}_{k,\ell})) = \mathrm{tr}(\Gamma_{k,\ell}^{-1/2}\widetilde{G}_{k,\ell}\widetilde{G}_{k,\ell}^T) = \left\langle \widetilde{G}_{k,\ell}, Z_{k,\ell}\right\rangle = \|Z_{k,\ell}\|_E \left\langle \widetilde{G}_{k,\ell}, \mathcal{N}_\ell(Z_{k,\ell})\right\rangle,$$

proving (3.1). Moreover, Assumption 4 also holds on block ℓ. Indeed, for any G_ℓ, the second and last part of (2.8) give

$$\mathrm{tr}(\mathcal{U}_{k,\ell}(\widetilde{G}_\ell)) = \mathrm{tr}(\widetilde{G}_\ell\widetilde{G}_\ell^T) = \|\widetilde{G}_\ell\|_E^2 = \|\widetilde{G}_\ell\|_{*,\ell}^2$$

so that (3.3) holds with $\kappa_\circ = 1$.

Thus, whenever a block ℓ is endowed with this full matrix geometry, Theorem 3.10, and Corollary 3.11 apply to that block without modification.

4.3 Diagonal AdaGrad [15, 39]

With the choices (2.9), we have that

$$\mathrm{tr}(\Gamma_{k,\ell}^{-1}\mathcal{U}_{k,\ell}(\widetilde{G}_{k,\ell})) = \sum_{i=1}^{n_\ell} \frac{[\widetilde{G}_{k,\ell}]_i^2}{[\Gamma_{k,\ell}]_{ii}} = \|Z_{k,\ell}\|_E^2,$$

which proves (3.2). Moreover,

$$\mathrm{tr}(\Gamma_{k,\ell}^{-1/2}\mathcal{U}_{k,\ell}(\widetilde{G}_{k,\ell})) = \sum_{i=1}^{n_\ell} \frac{[\widetilde{G}_{k,\ell}]_i^2}{\sqrt{[\Gamma_{k,\ell}]_{ii}}} = \left\langle \widetilde{G}_{k,\ell}, Z_{k,\ell}\right\rangle = \|Z_{k,\ell}\|_E \left\langle \widetilde{G}_{k,\ell}, \mathcal{N}_\ell(Z_{k,\ell})\right\rangle,$$

which proves (3.1). Moreover, Assumption 4 also holds in this case. Indeed, for any $G_\ell \in \mathbb{R}^{n_\ell}$,

$$\mathcal{U}_{k,\ell}(G_\ell) = \mathrm{diag}\big([G_\ell]_i^2\big),$$

and therefore

$$\mathrm{tr}(\mathcal{U}_{k,\ell}(G_\ell)) = \sum_{i=1}^{n_\ell} [G_\ell]_i^2 = \|G_\ell\|_E^2 = \|G_\ell\|_{*,\ell}^2,$$

so that (3.3) holds with $\kappa_\circ = 1$.

Thus, as above, Theorem 3.10 and Corollary 3.11 apply without modification to this diagonal AdaGrad geometry on block ℓ.

4.4 Adaptive MUON/AdaGO [28, 44, 56]

Consider the choices (2.10) and (2.11). For $\ell \in \mathcal{A}$, the verification of (3.1), (3.2) and (3.3) is identical to that of AdaNorm. For $\ell \in \mathcal{M}$, we have

$$\mathrm{tr}(\Gamma_{k,\ell}^{-1}\mathcal{U}_{k,\ell}(\widetilde{G}_{k,\ell})) = \frac{\|\widetilde{G}_{k,\ell}\|_*^2}{\gamma_{k,\ell}} = \|Z_{k,\ell}\|_*^2,$$

which proves (3.2). Therefore,

$$\|Z_{k,\ell}\|_* \left\langle \widetilde{G}_{k,\ell}, \mathcal{N}_\ell(Z_{k,\ell}) \right\rangle = \frac{\|\widetilde{G}_{k,\ell}\|_*^2}{\sqrt{\gamma_{k,\ell}}} = \mathrm{tr}(\Gamma_{k,\ell}^{-1/2}\mathcal{U}_{k,\ell}(\widetilde{G}_{k,\ell})),$$

which proves (3.1). Moreover, Assumption 4 also holds in this case. Indeed, for any block ℓ and any $\widetilde{G}_\ell$,

$$\mathrm{tr}(\mathcal{U}_{k,\ell}(\widetilde{G}_\ell)) = \begin{cases} \mathrm{tr}\left(\frac{\|\widetilde{G}_\ell\|_*^2}{d_\ell} I_{d_\ell}\right) = \|\widetilde{G}_\ell\|_*^2 & \text{if} \quad \ell \in \mathcal{M}, \\ \mathrm{tr}\left(\frac{\|\widetilde{G}_\ell\|_E^2}{d_\ell} I_{d_\ell}\right) = \|\widetilde{G}_\ell\|_E^2 = \|\widetilde{G}_\ell\|_*^2 & \text{if} \quad \ell \in \mathcal{A}. \end{cases}$$

Thus (3.3) holds with $\kappa_\circ = 1$ for all $\ell \in \{1, \ldots, L\}$. Therefore Theorem 3.10 and Corollary 3.11 apply without modification to this adaptive MUON geometry.

5 Conclusions

We have provided a general framework for first-order optimization algorithms using adaptively preconditioned gradients but no function values. We have shown that this framework covers a few of the popular adaptive first-order methods, including AdaGrad, full AdaGrad, AdaNorm and adaptive Muon, as well as any Cartesian mix of these. We have provided a fully stochastic rate-of-convergence analysis for all methods in the framework, with ¡and without momentum, using reasonable (although necessarily stonger than affine variance) assumptions on the variance of the gradient oracle and without assuming bounded stochastic gradients or small enough stepsizes.

Our results open several theoretical research issues. We anticipate that, as is the case for AdaGrad, the framework can be adapted to handle bounds [6] or more general convex [25, 13, 1] or nonconvex [16, 7, 22, 50] constraints, as well as second-order information (should it be available), but this requires verifications that are beyond the scope of the present paper. Other questions also remain. Can the theory be extended to cover other methods, like RMSProp [46] or Shampoo [25]? Is approximate normalization (i.e. approximate $\mathcal{N}_\ell(\cdot)$) acceptable? Can the generality of the preconditioner update be exploited to derive new preconditioners? Does the Cartesian framework allow more general algorithmic settings? Can the freedom to choose a different preconditioner update at each iteration be used to advantage? All these topics are the subject of ongoing investigation.

Acknowledgments

The authors are much indebted to two committed and timely referees for their patience with an imperfect manuscript and their many helpful and perceptive suggestions. Philippe Toint is also grateful for the continued friendly support of the APO team at ENSEEIHT (Toulouse).

Conflict of Interest

The authors confirm that there is no conflict of interest related to this paper.

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 *Proceedings of the 35th Conference on Neural Information Processing Systems (NEURIPS 2021)*, 2021.

[3] R. Anil, V. Gupta, T. Koren, K. Regan, and Y. Singer. Scalable second order optimization for deep learning. In *Proceedings of the 37th International Conference on Machine Learning (ICML)*, volume 119 of *Proceedings of Machine Learning Research*, pages 459–469. PMLR, 2020.

[4] A. Attia and T. Koren. SGD with AdaGrad stepsizes: Full adaptivity with high probability to unknown parameters, unbounded gradients and affine variance. arXiv:2302.08783, 2023.

[5] L. Balles, F. Pedregosa, and N. Le Roux. The geometry of sign gradient descent. arXiv:2002.08056, 2020.

[6] S. Bellavia, G. Gratton, B. Morini, and Ph. L. Toint. Fast stochastic Adagrad for nonconvex bound-constrained optimization. arXiv:2505:06374, 2025.

[7] S. Bellavia, G. Gratton, B. Morini, and Ph. L. Toint. An objective-function-free algorithm for general smooth constrained optimization. arXiv:2602:11770, 2026.

[8] J. Bernstein and L. Newhouse. Modular duality in deep learning. arXiv:2410.21265v2, 2024.

[9] J. Bernstein and L. Newhouse. Old optimizer, new norm: an anthology. arXiv:2409.20325v2, 2024.

[10] R. Bhatia. *Matrix Analysis*. Number 169 in Graduate Texts in Mathematics. Springer Verlag, Heidelberg, Berlin, New York, 1996.

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

[12] S. Boyd and L. Vandenberghe. *Convex Optimization*. Cambridge University Press, Cambridge, England, 2004.

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

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

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

[16] Y. Fang, S. Na, M. Mahoney, and M. Kolar. Fully stochastic trust-region sequential quadratic programming for equality constrained optimization problems. *SIAM Journal on Optimization*, 34(2):2007–2037, 2024.

[17] M. Faw, L. Rout, C. Caramanis, and S. Shakkottai. Beyond uniform smoothness: A stopped analysis of adaptive SGD. arXiv:2302.06570, 2023.

[18] 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 *Proceedings of 35th Conference on Learning Theory*, volume 178, pages 313–355, 2022.

[19] Th. Flynn. The duality structure gradient descent algorithm: analysis and application to neural networks. arXiv:1708.00523v8, 2017.

[20] B. Ginsburg, P. Castonguay, O. Hrinchuk, O. Kuchaiev, R. Leary, V. Lavrukhin, J. Li, H. Nguyen, Y. Zhang, and J. M. Cohen. Training deep networks with stochastic gradient normalized by layerwise adaptive second moments. arXiv:1905.11286v3, 2019.

[21] S. Gratton, S. Jerad, and Ph. L. Toint. Complexity and performance for two classes of noise-tolerant first-order algorithms. *Optimization Methods and Software*, 40(6):1473–1499, 2025.

[22] S. Gratton and Ph. L. Toint. An objective-function-free algorithm for nonconvex stochastic optimization with deterministic equality and inequality constraints. arXiv:2603.29685, 2026.

[23] Z. Guo, W. Yin, R. Jin, and T. Yang. A novel convergence analysis for algorithms of the Adam family. In *Proceedings of OPT2021: 13th Annual Workshop on Optimization for Machine Learning*, 2021.

[24] V. Gupta, T. Koren, and Y. Singer. A unified approach to adaptive regularization in online and stochastic optimization. arXiv:1706.06569, 2017.

[25] V. Gupta, T. Koren, and Y. Singer. Shampoo: Preconditioned stochastic tensor optimization. In *Proceedings of the International Conference on Machine Learning*, pages 1842–1850, 2018.

[26] Y. Hong and J. Lin. Revisting convergence of Adagrad with relaxed assumptions. arXiv:2403.13794v2, 2024.

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

[28] K. Jordan, Y. Jin, V. Boza, Y. Jiacheng, F. Cecista, L. Newhouse, and J. Bernstein. URL:https://kellerjordan.github.io.posts/muon, 2024.

[29] 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.

[30] D. Kingma and J. Ba. Adam: A method for stochastic optimization. In *Proceedings in the International Conference on Learning Representations (ICLR)*, 2015.

[31] D. Kovalev. SGD with adaptive preconditioning: Unified analysis and momentum acceleration. arXiv:2506.23804, 2025.

[32] H. Li, Y. Dong, and Z Lin. On the $\mathcal{O}(\sqrt{d}/t^{1/4})$ convergence rate for RMSProp and its momentum extension measured in ℓ_1 norm. *Journal of Machine Learning Research*, 26:1–25, 2025.

[33] J. Li and M. Hong. A note on the convergence of Muon. arXiv:2502.02900, 2025.

[34] 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.

[35] L. Liu, Z. Xu, Z. Zhang, H. Kang, Z. Li, C. Liang, W. Chen, and T. Zhao. COSMOS: A hybrid adaptive optimizer for memory-efficient training of LLMs. arXiv:2502.17410, 2025.

[36] 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.

[37] K. Löwner. Über monotone Matrixfunktionen. *Mathematisch Zeitung*, 38:177–216, 1934.

[38] J. Martens and R. Grosse. Optimizing neural networks with Kronecker-factored approximate curvature. In *International Conference on Machine Learning*, 2015.

[39] B. McMahan and M. Streeter. Adaptive bound optimization for online convex optimization. In *Conference on Learning Theory*, page 244sq, 2010.

[40] M. C. Mukkamala and M. Hein. Variants of RMSProp and Adagrad with logarithmic regret bounds. In *Proceedings of the 34th International Conference on Machine Learning*, page 2545–2553, 2017.

[41] Th. Pethick, W. Xie, K. Antonakopoulos, Z. Zhu, A. Silveti-Falls, and V. Cevher. Training deep learning models with norm-constrained LMOs. arXiv:2502.07529v2, 2025.

[42] S. Reddi, S. Kale, and S. Kumar. On the convergence of Adam and beyond. In *Proceedings in the International Conference on Learning Representations (ICLR)*, 2018.

[43] H. Robbins and S. Monroe. A stochastic approximation method. *The Annals of Mathematical Statistics*, 22(3):400–407, 1951.

[44] C. Si, D. Zhang, and W. Shen. AdaMuon: Adaptive Muon optimizer. arXiv2507.11005, 2025.

[45] M. Streeter and B. McMahan. Less regret via online conditioning, 2010.

[46] T. Tieleman and G. Hinton. Lecture 6.5-RMSPROP. COURSERA: Neural Networks for Machine Learning, 2012.

[47] Ph. L. Toint. Divergence of the ADAM algorithm with fixed-stepsize: a (very) simple example. In K. Fackeldey, A. Kannan, S. Pokutta, K. Sharma, D. Walter, A. Walther, and M. Weiser, editors, *Mathematical Optimization for Machine Learning: Proceedings of the MATH+ Thematic Einstein Semester 2023*, pages 195–199. De Gruyter Brill, 2025.

[48] N. Vyas, D. Morwani, R. Zhao, M. Kwun, I. Shapira, B. Brandfonbrener, L. Janson, and S. Kakade. SOAP: Improving and stabilizing Shampoo using Adam. arXiv:2409.11321, 2024.

[49] B. Wang, H. Zhang, Z. Ma, and W. Chen. Convergence of Adagrad for non-convex objectives: simple proofs and relaxed assumptions. In *Proceedings of the Thirty-Sixth Conference on Learning Theory*, pages 161–190, 2023.

[50] Q. Wang, Ch. Peirmarini, Y. Zhu, and F. E. Curtis. Projected stochastic momentum methods for nonlinear equality-constrained optimization for machine learning. arXiv:2601.11785, 2026.

[51] R. Ward, X. Wu, and L. Bottou. Adagrad stepsizes: sharp convergence over nonconvex landscapes. In *Proceedings of the International Conference on Machine Learning (ICML2019)*, 2019.

[52] X. Wu, R. Ward, and L. Bottou. WNGRAD: Learn the learning rate in gradient descent. arXiv:1803.02865, 2018.

[53] N. Xiao, X. Hu, X. Lin, and K.-C. Toh. Adam family methods for nonsmooth optimization with convergence guarantees. *Journal of Machine Learning Research*, 25, 2024.

[54] S. Xie, T. Wang, S. Reddi, S. Kumar, and Z. Li. Structured preconditioners in adaptive optimization: A unified analysis. arXiv:2503.10537, 2025.

[55] A. W. Yu, L. Huang, Q. Lin, R. Salakhutdinov, and J. Carbonell. Block-normalized gradient method; an empirical study for training deep neural network. arXiv:1707.04822v2, 2017.

[56] M. Zhang, Y. Liu, and H. Schaeffer. AdaGrad meets Muon: Adaptive stepsizes for orthogonal updates. arXiv:2509.02981v2, 2025.

[57] Q. Zhang, Y. Zhou, and S. Zou. Convergence guarantees for RMSProp and Adam in generalized-smooth non-convex optimization with affine noise variance, 2025.

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

[59] Y. Zhang, C. Chen, N. Shi, R. Sun, and Z.-Q. Luo. Adam can converge without any modification on update rules. arXiv:2208.09632v5, 2022.

[60] F. Zou, L. Shen, Z. Jie, J. Sun, and W Liu. A sufficient condition for convergences of Adam and RMSProp. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2019.

Proofs for Theorem 3.14 and Corollary 3.15

The definition (2.4) immediately implies the following bounds.

Lemma A.1 We have that, for all $k \geq 0$ and all $\ell \in \{1, \ldots, L\}$, $\Gamma_{k,\ell}$ is symmetric positive definite and

$$\Gamma_{k,\ell}^{-1/2} \preceq \frac{1}{\sqrt{\varsigma}} I_\ell \;\text{ and }\; \Gamma_{k,\ell}^{-1} \preceq \frac{1}{\varsigma} I_\ell \tag{A.1}$$

Proof. Directly result from (2.4), the definition of $\mathcal{U}_{k,\ell}$ and $\Gamma_{k,\ell} = \Gamma_{-1,\ell} + \sum_{j=0}^{k} \mathcal{U}_{j,\ell}(\widetilde{G}_{j,\ell}) \succeq \Gamma_{-1,\ell} = \varsigma I_\ell$. □

We now consider the impact of using momentum (i.e. $\mu > 0$ and $Z_{k,\ell}$ is a multiple of $M_{k,\ell}$) on the structural identities of Assumption 5.

Lemma A.2 Suppose that $\mu > 0$ and that Assumptions 2 and 5 and (3.67) hold. Then

$$\|Z_{k,\ell}\|_{*,\ell} \left\langle \widetilde{G}_{k,\ell}, \mathcal{N}_\ell(Z_{k,\ell}) \right\rangle_F \geq \frac{1}{\kappa_\square} \Big[\mathrm{tr}\big(\Gamma_{k,\ell}^{-1/2} \mathcal{U}_{k,\ell}(\widetilde{G}_{k,\ell})\big) - \mathrm{tr}\big(\Gamma_{k,\ell}^{-1/2} \mathcal{U}_{k,\ell}(E_{k,\ell})\big) \Big] - \|Z_{k,\ell}\|_{*,\ell} \|E_{k,\ell}\|_{*,\ell} \tag{A.2}$$

and

$$\|Z_{k,\ell}\|_{*,\ell}^2 \leq \kappa_\square \mathrm{tr}\big(\Gamma_{k,\ell}^{-1} \mathcal{U}_{k,\ell}(\widetilde{G}_{k,\ell})\big) + \kappa_\square \,\mathrm{tr}\big(\Gamma_{k,\ell}^{-1} \mathcal{U}_{k,\ell}(E_{k,\ell})\big) \tag{A.3}$$

Proof. Using (3.52), the linearity of the inner product and (3.56), we have that

$$\begin{aligned} \|Z_{k,\ell}\|_{*,\ell} \left\langle \widetilde{G}_{k,\ell}, \mathcal{N}_\ell(Z_{k,\ell}) \right\rangle_F &= \|Z_{k,\ell}\|_{*,\ell} \left\langle M_{k,\ell}, \mathcal{N}_\ell(Z_{k,\ell}) \right\rangle_F - \|Z_{k,\ell}\|_{*,\ell} \left\langle E_{k,\ell}, \mathcal{N}_\ell(Z_{k,\ell}) \right\rangle_F \\ &= \mathrm{tr}\big(\Gamma_{k,\ell}^{-1/2} \mathcal{U}_{k,\ell}(M_{k,\ell})\big) - \|Z_{k,\ell}\|_{*,\ell} \left\langle E_{k,\ell}, \mathcal{N}_\ell(Z_{k,\ell}) \right\rangle_F . \end{aligned} \tag{A.4}$$

Now observe that (3.67) and (3.52) give that

$$\mathcal{U}_{k,\ell}(\widetilde{G}_{k,\ell}) = \mathcal{U}_{k,\ell}(M_{k,\ell} - E_{k,\ell}) \preceq \kappa_\square \mathcal{U}_{k,\ell}(M_{k,\ell}) + \kappa_\square \mathcal{U}_{k,\ell}(E_{k,\ell})$$

Because $\Gamma_{k,\ell}^{-1/2}$ is positive definite (Lemma A.1) and the trace is monotone for the Löwner semi-order, we have that

$$\mathrm{tr}\big(\Gamma_{k,\ell}^{-1/2} \mathcal{U}_{k,\ell}(\widetilde{G}_{k,\ell})\big) \leq \kappa_\square \,\mathrm{tr}\big(\Gamma_{k,\ell}^{-1/2} \mathcal{U}_{k,\ell}(M_{k,\ell})^2 2\big) + \kappa_\square \,\mathrm{tr}\big(\Gamma_{k,\ell}^{-1/2} \mathcal{U}_{k,\ell}(E_{k,\ell})\big).$$

Substituting the resulting lower bound on $\Gamma_{k,\ell}^{-1/2} \mathcal{U}_{k,\ell}(M_{k,\ell})$ into (A.4), we deduce that

$$\begin{aligned} \|Z_{k,\ell}\|_{*,\ell} \left\langle \widetilde{G}_{k,\ell}, \mathcal{N}_\ell(Z_{k,\ell}) \right\rangle_F \geq \frac{1}{\kappa_\square} &\Big[\mathrm{tr}\big(\Gamma_{k,\ell}^{-1/2} \mathcal{U}_{k,\ell}(\widetilde{G}_{k,\ell})\big) - \mathrm{tr}\big(\Gamma_{k,\ell}^{-1/2} \mathcal{U}_{k,\ell}(E_{k,\ell})\big) \Big] \\ &- \|Z_{k,\ell}\|_{*,\ell} \left\langle E_{k,\ell}, \mathcal{N}_\ell(Z_{k,\ell}) \right\rangle_F \end{aligned}$$

and (A.2) then results from the fact that, by duality of the primal and dual norms and (2.3),

$$\left\langle E_{k,\ell}, \mathcal{N}_\ell(Z_{k,\ell}) \right\rangle_F \leq \|E_{k,\ell}\|_{*,\ell} \|\mathcal{N}_\ell(Z_{k,\ell})\|_\ell = \|E_{k,\ell}\|_{*,\ell}.$$

In order to prove (A.3), we first note that (3.52) and (3.67) yield that

$$\mathcal{U}_{k,\ell}(M_{k,\ell}) = \mathcal{U}_{k,\ell}(\widetilde{G}_{k,\ell} + E_{k,\ell}) \preceq \kappa_\square \mathcal{U}_{k,\ell}(\widetilde{G}_{k,\ell}) + \kappa_\square \mathcal{U}_{k,\ell}(E_{k,\ell})$$

and thus, again using the monotonicity of the trace for the Löwner semi-order, that

$$\mathrm{tr}\big(\Gamma_{k,\ell}^{-1}\mathcal{U}_{k,\ell}(M_{k,\ell})\big) \leq \kappa_\square\,\mathrm{tr}\big(\Gamma_{k,\ell}^{-1}\mathcal{U}_{k,\ell}(\widetilde{G}_{k,\ell})\big) + \kappa_\square\,\mathrm{tr}\big(\Gamma_{k,\ell}^{-1}\mathcal{U}_{k,\ell}(E_{k,\ell})\big).$$

Substituting this inequality in (3.57) then gives (A.3). □

Observe that (A.2) is (up to a constant factor) a perturbation of a version of (3.56) for block ℓ at iteration k, the perturbation term being $\mathrm{tr}\big(\Gamma_{k,\ell}^{-1/2}\mathcal{U}_{k,\ell}(E_{k,\ell})\big) + \|Z_{k,\ell}\|_{*,\ell}\|E_{k,\ell}\|_{*\ell}$. The same is true for (A.3), in which case the perturbation of (3.57) is $\mathrm{tr}\big(\Gamma_{k,\ell}^{-1}\mathcal{U}_{k,\ell}(E_{k,\ell})\big)$. We now show that the terms occurring in these perturbations can be bounded.

Lemma A.3 Suppose that Assumption 2 and 5 and (3.67) hold. Then

$$\sum_{j=0}^{k}\sum_{\ell=1}^{L}\mathbb{E}[\|Z_{j,\ell}\|_{*,\ell}\|E_{j,\ell}\|_{*,\ell}] \leq \frac{3\theta_k^2}{(1-\mu_{\max})^2} + \frac{3L_G^2\eta^2}{2(1-\mu_{\max})^2}\sum_{j=0}^{k}\bar{\mu}_j^2\mathbb{E}\big[\|Z_j\|_*^2\big] + 2\sum_{j=0}^{k}\mathbb{E}\big[\|Z_j\|_*^2\big] \tag{A.5}$$

$$\sum_{j=0}^{k}\sum_{\ell=1}^{L}\mathbb{E}\Big[\mathrm{tr}\big(\Gamma_{j,\ell}^{-1/2}\mathcal{U}_{k,\ell}(E_{j,\ell})\big)\Big] \leq \frac{6\kappa_\diamond\theta_k^2}{(1-\mu_{\max})^2\sqrt{\varsigma}} + \frac{3\kappa_\diamond L_G^2\eta^2}{(1-\mu_{\max})^2\sqrt{\varsigma}}\sum_{j=0}^{k}\bar{\mu}_j^2\mathbb{E}\big[\|Z_j\|_*^2\big] \tag{A.6}$$

and

$$\sum_{j=0}^{k}\sum_{\ell=1}^{L}\mathbb{E}\Big[\mathrm{tr}\big(\Gamma_{k,\ell}^{-1}\mathcal{U}_{k,\ell}(E_{k,\ell})\big)\Big] \leq \frac{6\kappa_\diamond\theta_k^2}{(1-\mu_{\max})^2\varsigma} + \frac{3\kappa_\diamond L_G^2\eta^2}{(1-\mu_{\max})^2\varsigma}\sum_{j=0}^{k}\bar{\mu}_j^2\mathbb{E}\big[\|Z_j\|_*^2\big]\,. \tag{A.7}$$

Proof. For each block $\ell \in \{1,\ldots,L\}$ and each iteration $j \in \{0,\ldots,k\}$, we have, using $ab \leq (a^2+b^2)/2$, that

$$\|Z_{k,\ell}\|_{*,\ell}\|E_{k,\ell}\|_{*,\ell} \leq \frac{1}{2}\|Z_{k,\ell}\|_{*,\ell}^2 + \frac{1}{2}\|E_{k,\ell}\|_{*,\ell}^2$$

and thus

$$\sum_{j=0}^{k}\sum_{\ell=1}^{L}\mathbb{E}[\|Z_{j,\ell}\|_{*,\ell}\|E_{j,\ell}\|_{*,\ell}] \leq \frac{1}{2}\sum_{j=0}^{k}\sum_{\ell=1}^{L}\mathbb{E}\big[\|Z_{j,\ell}\|_{*,\ell}^2\big] + \frac{1}{2}\sum_{j=0}^{k}\sum_{\ell=1}^{L}\mathbb{E}\big[\|E_{j,\ell}\|_{*,\ell}^2\big]$$

Substituting now (3.53) in this inequality gives (A.5). Consider now the trace terms. Because of (A.1), the monotonicity of the trace for the Löwner semi-order and Assumption 3.67, we have that

$$\mathrm{tr}\big(\Gamma_{k,\ell}^{-1/2}\mathcal{U}_{k,\ell}(E_{k,\ell})\big) \leq \frac{1}{\sqrt{\varsigma}}\,\mathrm{tr}\big(\mathcal{U}_{k,\ell}(E_{k,\ell})\big) \leq \frac{\kappa_\diamond}{\sqrt{\varsigma}}\|E_{k,\ell}\|_{*,\ell}^2$$

Taking full expectation, summing for $j \in \{0,\ldots,k\}$ and $\ell \in \{1,\ldots,L\}$ and substituting (3.53) then gives (A.6). The proof of (A.7) is entirely similar, using ς instead of $\sqrt{\varsigma}$. □

We may now substitute the bounds of Lemma A.3 into those of Lemma A.2.

Lemma A.4 Suppose that $\mu > 0$ and that Assumption 2 and 5 and (3.67) hold. Suppose in addition that (3.68) is satisfied. Then

$$\sum_{j=0}^{k}\sum_{\ell=1}^{L}\mathbb{E}\Big[\|Z_{j,\ell}\|_{*,\ell}\,\Big\langle\widetilde{G}_{j,\ell},\mathcal{N}_\ell(Z_{j,\ell})\Big\rangle_F\Big] \geq \frac{1}{\kappa_\square}\sum_{j=0}^{k}\sum_{\ell=1}^{L}\mathbb{E}\Big[\mathrm{tr}\big(\Gamma_{j,\ell}^{-1/2}\mathcal{U}_{j,\ell}(\widetilde{G}_{k,\ell})\big)\Big] - \kappa_{1,\theta}\,\theta_k^2 - \kappa_{1,z}\sum_{j=0}^{k}\mathbb{E}\big[\|Z_j\|_*^2\big] \tag{A.8}$$

with

$$\kappa_{1,\theta} = \frac{6\kappa_\diamond}{(1-\mu_{\max})^2\sqrt{\varsigma}} + \frac{3}{(1-\mu_{\max})^2} \;\text{ and }\; \kappa_{1,z} = \frac{3\kappa_\diamond\mu_{\max}^2L_G^2\eta^2}{(1-\mu_{\max})^2\sqrt{\varsigma}} + \frac{3\mu_{\max}^2L_G^2\eta^2}{2(1-\mu_{\max})^2} + 2, \tag{A.9}$$

and

$$\sum_{j=0}^{k}\mathbb{E}\big[\|Z_j\|_*^2\big] \leq 2\kappa_\square\sum_{j=0}^{k}\sum_{\ell=1}^{L}\mathbb{E}\Big[\mathrm{tr}\big(\Gamma_{j,\ell}^{-1}\mathcal{U}_{j,\ell}(\widetilde{G}_{j,\ell})\big)\Big] + \kappa_{2,\theta}\,\theta_k^2 + \kappa_{2,z}, \tag{A.10}$$

where

$$\kappa_{2,\theta} = \frac{12\kappa_\square\kappa_\diamond}{(1-\mu_{\max})^2\varsigma} \quad\text{and}\quad \kappa_{2,z} = \begin{cases} 0 & \text{if the first part of (3.68) holds,} \\ \frac{3\kappa_\square\kappa_\diamond L_G^2\eta^2}{(1-\mu_{\max})^2\varsigma}\,\kappa_{\mu Z} & \text{if the second part of (3.68) holds.} \end{cases} \tag{A.11}$$

Proof. The inequality (A.8) is obtained by substituting (A.5) and (A.6) into (A.2), Moreover, taking the expectation in (A.3), summing for $j \in \{0,\ldots,k\}$ and $\ell \in \{1,\ldots,L\}$ and substituting (A.7) gives that

$$\sum_{j=0}^{k}\sum_{\ell=1}^{L}\mathbb{E}\big[\|Z_{k,\ell}\|_{*,\ell}^2\big] \leq \kappa_\square\sum_{j=0}^{k}\sum_{\ell=1}^{L}\mathbb{E}\Big[\mathrm{tr}\big(\Gamma_{k,\ell}^{-1}\mathcal{U}_{k,\ell}(\widetilde{G}_{j,\ell})\big)\Big] + \frac{6\kappa_\square\kappa_\diamond}{(1-\mu_{\max})^2\varsigma}\,\theta_k^2 + \frac{3\kappa_\square\kappa_\diamond L_G^2\eta^2}{(1-\mu_{\max})^2\varsigma}\sum_{j=0}^{k}\bar{\mu}_j^2\mathbb{E}\big[\|Z_j\|_*^2\big].$$

Suppose first that the first part of (3.68) holds. Then, since $\bar{\mu}_j \leq \mu_{\max}$,

$$\frac{3\kappa_\square\kappa_\diamond\mu_{\max}^2L_G^2\eta^2}{(1-\mu_{\max})^2\varsigma} \leq \frac{1}{2}$$

and (A.10) follows with

$$\kappa_{2,z} = 0 \;\text{ and }\; \kappa_{2,\theta} = \frac{12\kappa_\square\kappa_\diamond}{(1-\mu_{\max})^2\varsigma}.$$

Alternatively, if the second part of (3.68) holds, then (A.10) follows with

$$\kappa_{2,z} = \frac{3\kappa_\square\kappa_\diamond L_G^2\eta^2}{(1-\mu_{\max})^2\varsigma}\,\kappa_{\mu Z} \;\text{ and }\; \kappa_{2,\theta} = \frac{6\kappa_\square\kappa_\diamond}{(1-\mu_{\max})^2\varsigma}.$$

□

From this point on, the analysis follows the lines of the argument of Section 3 with Lemma A.4 providing an alternate set of structural inequalities. Lemmas 3.2 to 3.4 are unchanged. The modifications to Theorem 3.5 are minor. It is restated as follows.

Theorem A.5 Suppose that Assumptions 1, 2 and 5, (3.13) and (3.68) hold. Define Δ_k as in (3.14). Then, for every $k \geq 0$,

$$\mathbb{E}\left[\sum_{\ell=1}^{L} \mathrm{tr}(\Gamma_{k,\ell}^{1/2})\right] \leq \kappa_{\rm gap} + \varepsilon(\nu_k, \theta_k) + \sqrt{2\kappa_\square}\,\kappa_\square\,\nu_k\sqrt{\Delta_k} + \kappa_\Delta\kappa_\square^2\Delta_k, \tag{A.12}$$

where

$$\kappa_{\rm gap} \stackrel{\rm def}{=} \kappa_\square\,\frac{f(X_0)-f_{\rm low}}{\eta} + \sqrt{\varsigma}\,N + \kappa_\square\left(\kappa_{1,z} + \omega + \frac{L_G\eta}{2}\right)\kappa_{2,z}, \tag{A.13}$$

$$\varepsilon(\nu,\theta) = \kappa_\square\left(\sqrt{\kappa_{2,z}}\,\nu + \sqrt{\kappa_{2,\theta}}\,\nu\theta + \kappa_{\theta\theta}\theta^2\right) \tag{A.14}$$

and

$$\kappa_{\theta\theta} \stackrel{\rm def}{=} \kappa_{1,\theta} + \left(\kappa_{1,z} + \omega + \frac{L_G\eta}{2}\right)\kappa_{2,\theta}, \quad \kappa_\Delta \stackrel{\rm def}{=} 2\left(\kappa_{1,z} + \omega + \frac{L_G\eta}{2}\right). \tag{A.15}$$

Proof. Inequality (A.8) gives, in place of (3.18),

$$\eta\sum_{j=0}^{k}\sum_{\ell=1}^{L}\mathbb{E}\Big[\|Z_{j,\ell}\|_{*,\ell}\left\langle\widetilde{G}_{j,\ell}\right\rangle\mathcal{N}_\ell(Z_{j,\ell})\Big] \geq \frac{\eta}{\kappa_\square}\mathbb{E}\left[\sum_{\ell=1}^{L}\mathrm{tr}(\Gamma_{k,\ell}^{1/2}) - \sum_{\ell=1}^{L}\mathrm{tr}(\Gamma_{-1,\ell}^{1/2})\right] - \eta\kappa_{1,\theta}\theta_k^2 - \eta\kappa_{1,z}\zeta_k,$$

where

$$\zeta_k \stackrel{\rm def}{=} \sum_{j=0}^{k}\sum_{\ell=1}^{L}\mathbb{E}\big[\|Z_{j,\ell}\|_{*,\ell}^2\big]\,.$$

Proceeding as in the proof of Theorem 3.5, while keeping the stochastic error $\widetilde{G}_{j,\ell} - G_{j,\ell}$ separate, we obtain

$$\begin{aligned}\frac{\eta}{\kappa_\square}&\mathbb{E}\left[\sum_{\ell=1}^{L}\mathrm{tr}(\Gamma_{k,\ell}^{1/2}) - \sum_{\ell=1}^{L}\mathrm{tr}(\Gamma_{-1,\ell}^{1/2})\right]\\ &\leq f(X_0) - f_{\rm low} + \eta\kappa_{1,\theta}\theta_k^2 + \eta\sum_{j=0}^{k}\sum_{\ell=1}^{L}\mathbb{E}\Big[\|Z_{j,\ell}\|_{*,\ell}\|\widetilde{G}_{j,\ell} - G_{j,\ell}\|_{*,\ell}\Big] + \left(\eta\kappa_{1,z} + \frac{L_G\eta^2}{2}\right)\zeta_k.\end{aligned}$$

By the Cauchy–Schwarz inequality and (3.13),

$$\begin{aligned}\sum_{j=0}^{k}\sum_{\ell=1}^{L}\mathbb{E}\Big[\|Z_{j,\ell}\|_{*,\ell}\|\widetilde{G}_{j,\ell} - G_{j,\ell}\|_{*,\ell}\Big] &\leq \sqrt{\sum_{j=0}^{k}\sum_{\ell=1}^{L}\mathbb{E}\Big[\|Z_{j,\ell}\|_{*,\ell}^2\Big]}\sqrt{\sum_{j=0}^{k}\sum_{\ell=1}^{L}\mathbb{E}\Big[\|\widetilde{G}_{j,\ell} - G_{j,\ell}\|_{*,\ell}^2\Big]}\\ &\leq \sqrt{\zeta_k}\,\sqrt{\nu_k^2 + \omega^2\zeta_k}\\ &\leq \nu_k\sqrt{\zeta_k} + \omega\zeta_k.\end{aligned}$$

Hence

$$\frac{\eta}{\kappa_\square}\mathbb{E}\left[\sum_{\ell=1}^{L}\mathrm{tr}(\Gamma_{k,\ell}^{1/2}) - \sum_{\ell=1}^{L}\mathrm{tr}(\Gamma_{-1,\ell}^{1/2})\right] \leq f(X_0) - f_{\rm low} + \eta\kappa_{1,\theta}\theta_k^2 + \eta\nu_k\sqrt{\zeta_k} + \eta\left(\kappa_{1,z} + \omega + \frac{L_G\eta}{2}\right)\zeta_k.$$

On the other hand, the definition of ζ_k, (A.10), (3.11) and (3.14) give

$$\zeta_k \leq \kappa_{2,z} + \kappa_{2,\theta}\theta_k^2 + 2\kappa_\square\sum_{j=0}^{k}\sum_{\ell=1}^{L}\mathbb{E}\Big[\mathrm{tr}\Big(\Gamma_{j,\ell}^{-1}\mathcal{U}_{j,\ell}(\widetilde{G}_{j,\ell})\Big)\Big] \leq \kappa_{2,z} + \kappa_{2,\theta}\theta_k^2 + 2\kappa_\square\Delta_k.$$

Consequently, using $\sqrt{a+b+c} \leq \sqrt{a} + \sqrt{b} + \sqrt{c}$,

$$\sqrt{\zeta_k} \leq \sqrt{\kappa_{2,z}} + \sqrt{\kappa_{2,\theta}}\,\theta_k + \sqrt{2\kappa_\square \Delta_k}.$$

Substitution in the preceding inequality yields

$$\begin{aligned}\frac{\eta}{\kappa_\square}\mathbb{E}\left[\sum_{\ell=1}^{L}\mathrm{tr}(\Gamma_{k,\ell}^{1/2}) - \sum_{\ell=1}^{L}\mathrm{tr}(\Gamma_{-1,\ell}^{1/2})\right] &\leq f(X_0) - f_{\rm low} + \eta\nu_k\sqrt{\kappa_{2,z}} + \eta\nu_k\sqrt{\kappa_{2,\theta}}\,\theta_k \\ &\quad + \eta\left[\kappa_{1,\theta} + \left(\kappa_{1,z} + \omega + \frac{L_G\eta}{2}\right)\kappa_{2,\theta}\right]\theta_k^2 \\ &\quad + \sqrt{2\kappa_\square}\eta\nu_k\sqrt{\Delta_k} + 2\eta\kappa_\square\left(\kappa_{1,z} + \omega + \frac{L_G\eta}{2}\right)\Delta_k \\ &\quad + \eta\left(\kappa_{1,z} + \omega + \frac{L_G\eta}{2}\right)\kappa_{2,z}.\end{aligned}$$

Finally, since $\Gamma_{-1,\ell} = \varsigma I_\ell$ for all $\ell \in \{1,\ldots,L\}$,

$$\sum_{\ell=1}^{L}\mathrm{tr}(\Gamma_{-1,\ell}^{1/2}) = \sqrt{\varsigma}\,N,$$

and the result follows from the definitions (A.13)–(A.15). □

Note that the form of bound (A.12) differs very little from that of (3.15): besides using different constants, (A.12) now involves the term $\varepsilon(\nu_k,\theta_k)$.

Lemmas 3.6 and 3.7 are again unmodified. Because of the similarity between (A.12) and (3.15), the proof of Theorem 3.14 is then identical to those of Theorems 3.9 and 3.10, except that the crucial inequality (3.33) in Theorem 3.9 now holds with (3.34) replaced by

$$a = \kappa_{\rm gap} + \varepsilon(\nu_k,\theta_k), \quad b = 2\kappa_\square\sqrt{2\kappa_\square}\,\sqrt{N} \quad \text{and} \quad c = 4\kappa_\Delta\kappa_\square^2 N.$$

Finally, the proof of Corollary 3.15 is similar to that of Corollary 3.11. When considering ADPREC.MG rather than ADPREC, we see from (3.69) that now

$$\Theta_k = \mathcal{O}\left(\nu_k \max\left[\theta_k, \sqrt{\log(\nu_k)}\right]\right) \tag{A.16}$$

instead of (3.45), and we may derive (as done in Corollary 3.11 for (3.46)) that

$$\theta_k \leq \begin{cases} \dfrac{\sigma_{\rm tot}\mu_{\max}}{\sqrt{1-\alpha-2\beta}}(k+1)^{\frac{1}{2}(1-\alpha)-\beta} & \text{if} \quad \alpha+2\beta<1, \\ \sigma_{\rm tot}\mu_{\max}\sqrt{1+\log(k+1)} & \text{if} \quad \alpha+2\beta=1, \\ \sigma_{\rm tot}\mu_{\max}\zeta(\alpha+2\beta) & \text{if} \quad \alpha+2\beta>1. \end{cases} \tag{A.17}$$

In order to obtain (3.74) and given (3.46) and (A.17), we now verify which term in the maximum of (A.16) dominates for the different combinations of α and $\alpha+2\beta$.
• If $\alpha+2\beta<1$, then $\alpha<1$ and $\frac{1}{2}(1-\alpha)-\beta>0$, and thus

$$\max\left[(k+1)^{\frac{1}{2}(1-\alpha)-\beta}, \sqrt{\log\left((k+1)^{\frac{1}{2}(1-\alpha)}\right)}\right] = (k+1)^{\frac{1}{2}(1-\alpha)-\beta},$$

yielding $\Theta_k = \mathcal{O}((k+1)^{1-\alpha-\beta})$.
• If $\alpha<1$ and $\alpha+2\beta=1$,

$$\max\left[\sqrt{\log(k+1)}, \sqrt{\log\left((k+1)^{\frac{1}{2}(1-\alpha)}\right)}\right] = \sqrt{\log(k+1)},$$

giving $\Theta_k = \mathcal{O}(\sqrt{\log(k+1)}(k+1)^{\frac{1}{2}(1-\alpha)})$.
• If $\alpha < 1$ and $\alpha + 2\beta > 1$,

$$\max\left[\zeta(\alpha+2\beta), \sqrt{\log\left((k+1)^{\frac{1}{2}(1-\alpha)}\right)}\right] = \sqrt{\log\left((k+1)^{\frac{1}{2}(1-\alpha)}\right)} = \sqrt{\tfrac{1}{2}(1-\alpha)\log(k+1)},$$

and thus $\Theta_k = \mathcal{O}(\sqrt{\log(k+1)}(k+1)^{\frac{1}{2}(1-\alpha)})$.
• If $\alpha = 1$ and $\alpha + 2\beta = 1$,

$$\max\left[\sqrt{\log(k+1)}, \sqrt{\log\left(\sqrt{\log(k+1)}\right)}\right] = \sqrt{\log(k+1)},$$

implying $\Theta_k = \mathcal{O}(\log(k+1))$.
• If $\alpha = 1$ and $\alpha + 2\beta > 1$,

$$\max\left[\zeta(\alpha+2\beta), \sqrt{\log\left(\sqrt{\log(k+1)}\right)}\right] = \sqrt{\log\left(\sqrt{\log(k+1)}\right)} = \sqrt{\tfrac{1}{2}\log(\log(k+1))},$$

and $\Theta_k = \mathcal{O}(\sqrt{\log(k+1)}\sqrt{\log(\log(k+1))})$.
• If $\alpha > 1$ and $\alpha + 2\beta > 1$,

$$\max\left[\zeta(\alpha+2\beta), \sqrt{\log(\zeta(\alpha))}\right] = \mathcal{O}(1),$$

and $\Theta_k = \mathcal{O}(1)$.
Inserting these bounds on Θ_k in (3.72) then gives (3.74).

We conclude by noting that (3.56) and (3.57) are satisfied for all methods considered because the results from Section 4 still hold when the approximate gradient $\widetilde{G}_k$ is set to the momentum M_k. Moreover (3.67) can easily be verified for all methods, as replacing $\widetilde{G}_{k,\ell}$ by a general matrix U of suitable size in the derivations involving $\mathcal{U}_{k,\ell}(\cdot)$ implies that (3.67) holds with $\kappa_\square = 2$ and $\kappa_\diamond = 1$ for all these methods, except that $\kappa_\diamond = 1/N$ for AdaNorm.