Complexity and performance for two classes of noise-tolerant first-order algorithms

Gratton, Serge; Jerad, Sadok; Toint, Philippe

Published in:
Optimization Methods and Software

DOI:
10.1080/10556788.2025.2532736

Publication date:
2025

Complexity and performance for two classes of noise-tolerant first-order algorithms

S. Gratton[*], S. Jerad[†] and Ph. L. Toint[‡]

8 VII 2025

Abstract

Two classes of algorithms for optimization in the presence of noise are presented, that do not require the evaluation of the objective function. The first generalizes the well-known Adagrad method. Its complexity is then analyzed as a function of its parameters. A second class of algorithms is then derived whose complexity is at least as good as that of the first class. Initial numerical experiments on finite-sum problems arising from deep-learning applications suggest that methods of the second class may outperform those of the first.

Keywords: First-order methods, objective-function-free optimization, noisy gradients, Adagrad, convergence bounds, evaluation complexity.

1 Introduction

Minimization algorithms which can handle noisy evaluations of the objective function and/or gradients have generated a significant amount of research in the last few years [4, 5, 6, 8, 9, 10, 12, 15, 16, 18, 29, 30, 36, 38, 39, 41, 45, 46]. Interestingly, a number of these contributions [4, 5, 6, 8, 9, 10, 12, 15, 36] indicate that, when the (noisy) objective function is evaluated, its accuracy is significantly more critical to ensure convergence than that of the computed (noisy) derivatives. This may be the reason why methods where the problem is avoided by *not* evaluating the objective function (such as Adagrad [17], RMSProp [39], Adam [29] or AMSGrad [37]), have become very popular in the context of finite-sum minimization, where noise in the evaluation arises from sampling among a very large number of terms. That such methods can be provably convergent to first-order stationary points is quite remarkable, and the literature covering their theory is extensive. We now briefly survey some of the contributions most relevant in our context.

[*]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"

[†]ANITI, Université de Toulouse, INP, IRIT, Toulouse, France. Email: sadok.jerad@enseeiht.fr

[‡]NAXYS, University of Namur, Namur, Belgium. Email: philippe.toint@unamur.be. Partially supported by ANITI.

1.1 Related work

Several authors have been able to prove global convergence rates, including the recent contributions of [1, 16, 21, 19, 20, 28, 30, 40, 41, 45], where a global convergence analysis of the Adagrad method has been conducted under different assumptions. The paper [30] provides an analysis of a delayed Adagrad method which does not consider the sampled gradient at iteration k to compute its step size, and further variants that are arbitrarily close to Adagrad and require specific choices of hyperparameters. The first parameter-free analysis of Adagrad was proposed in [41] and later revisited by [16], where the dependence on some parameters was improved. In both cases the bounds were given in expectation. In [28], high-probability bounds are derived for a class of methods that includes Adagrad. In [45], complexity analyses of a large class of adaptive gradient methods (including Adagrad) are proposed, and improved convergence rates are proved under the "gradient sparsity" assumption of the gradient iteration sequence. Note that all of [16, 28, 41, 45] require the sampled gradient to be bounded. This requirement was circumvented in [1, 19, 20, 40] by allowing unbounded gradients boundedness using a new Lipschitz smoothness condition [44] and different noise assumptions, see [19, 40, 20, 1] for more details.

1.2 Our contributions

The present paper remains in the context of bounded gradients and extends some results of [41, 16] to achieve several goals.

1. The global rate of convergence result of [16] is shown to hold for an extended class of methods including the Adagrad algorithm.

2. Using the new analysis tools, a new class of methods is then proposed, whose global rate of convergence is shown to be very close to that of methods using (exact) function evaluations.

3. Numerical experiments with finite-sum problems arising from deep-learning applications indicate that method of the latter class may sometimes perform better than those of the former.

The presentation is organized as follows. A general framework of first-order trust-region algorithms is introduced in Section 2, in which two classes of algorithms (one of them containing the Adagrad method) are defined and analyzed (complexity-wise) in Sections 3 and 4, respectively. Numerical experiments in the finite-sum minimization context are presented in Section 5. Some conclusions are finally outlined in Section 6.

2 A first-order framework for minimizing noisy functions

We are interested in (approximately) solving the problem

$$\min_{x \in \mathbb{R}^n} F(x) \tag{2.1}$$

where F is a function from $\mathbb{R}^n$ to $\mathbb{R}$ contaminated by noise. Moreover, we assume that evaluating F at any given x to sufficient accuracy is either impossible or too costly. Evaluating a noisy gradient is however possible... and our only source of information about the problem.

While access to F or its exact gradient is impossible, we nevertheless make the following assumptions.

Assumption 2.1. The objective function $F(x)$ is continuously differentiable.

Assumption 2.2. Its exact gradient $G(x) \stackrel{\text{def}}{=} \nabla_x^1 f(x)$ is Lipschitz continuous with Lipschitz constant L, that is

$$\|G(x) - G(y)\| \leq L\|x - y\|$$

for all $x, y \in \mathbb{R}^n$.

Assumption 2.3. There exists a constant F_{low} such that, for all x, $F(x) \geq F_{\text{low}}$.

A standard consequence of Assumption 2.2 is that, for, any $x, s \in \mathbb{R}^n$,

$$F(x+s) \leq F(x) + G(x)^T s + \frac{L}{2}\|s\|^2 \tag{2.2}$$

(see Lemma 2.1 in [7] or Theorem A.8.3 in [14], for instance).
We now present a first-order *adaptively scaled gradient* algorithmic framework (ASGRAD), where, at iteration k, a *noisy* gradient $g_k = g(x_k)$ is evaluated and a step s_k defined that decreases the associated local linear model and whose size is determined by componentwise "scaling factors" $w_{i,k}$ to be chosen at each iteration. Our framework is formally described as follows.

Algorithm 2.1: The ASGRAD framework

Step 0: Initialization. x_0 and a constant $\gamma_{\text{low}} \in (0, 1]$ are given. Set $k = 0$.

Step 1: Step computation. Evaluate g_k and set

$$s_k = \gamma_k s_k^L, \tag{2.3}$$

with

$$s_{i,k}^L = -\frac{g_{i,k}}{w_{i,k}} \quad (i \in \{1, \ldots, n\}) \tag{2.4}$$

for a stepsize $\gamma_k \in [\gamma_{\text{low}}, 1]$ and positive scaling factors $w_{i,k}$.

Step 2: New iterate. Define

$$x_{k+1} = x_k + s_k, \tag{2.5}$$

increment k by one and return to Step 1.

We stress that g_k (as evaluated in Step 1) is a noisy random gradient evaluation. The algorithms of the ASGRAD framework therefore generate a stochastic process

$$\{x_k, g_k, \gamma_k, s_k^L, s_k\}$$

on some probability space $(\Omega, \mathcal{F}, \mathbb{P})$. The associated expectation operator will be denoted by $\mathbb{E}[\cdot]$ and $\mathbb{E}_k[\cdot]$ will stand for the conditional expectation knowing $\{g_0, \ldots, g_{k-1}\}$. All algorithms

in our framework may clearly be interpreted as variants of Stochastic Gradient Descent, allowing for a variety of stepsize (learning rate) rules.

We will, in what follows, assume that the noisy gradient g_k is a bounded non-biased estimator of the true gradient, that is

Assumption 2.4. We have that, for all $k \geq 0$, $\mathbb{E}_k[g_k] = G(x_k)$. Moreover, there exists a constant $\kappa_g \geq 1$ such that $\|g_k\|_\infty \leq \kappa_g$ for all $k \geq 0$ and all realizations of the algorithm. In addition, we assume that γ_k is mesurable with respect to $\{g_0, \ldots, g_{k-1}\}$.

This assumption that gradients are bounded maybe quite realistic in practice[(1)], for instance when the iterates remain in a compact subset of $\mathbb{R}^n$ and has been extensively used in the analysis of stochastic first-order methods (see [41, 16, 45, 43, 28]) immediately implies that

$$\|G(x_k)\|_\infty \leq \kappa_g \quad \text{for all} \quad k \geq 0. \tag{2.6}$$

The assumption on the stepsize γ_k is consistent with current best practices in training deep neural networks: to achieve state-of-the-art performance, the step size γ_k is often either set to a specific value at the beginning and divided by a fixed constant at (approximately) regular intervals [42] or periodically warm-restarted [33].
The reader has undoubtedly noted that we have not been very specific regarding how the scaling factors $w_{i,k}$ are selected, and a whole range of options is possible. This justifies our choice to consider ASGRAD as an *algorithmic framework*, covering many possible such choices. The rest of this paper is devoted to the analysis of two specific classes of interest.

3 An Adagrad-inspired class of ASGRAD algorithms

In the first considered ASGRAD class, the scaling factors are inspired by the definition of the Adagrad algorithm [17]. More specifically, we make the following additional assumptions.

Assumption 3.1. For each $i \in \{1, \ldots, n\}$ and $k \geq 0$, there exist a constant $\varsigma_i > 0$ and a random variable $v_{i,k}$ such that $v_{i,k} \geq \varsigma_i$ and $w_{i,k} = (v_{i,k})^\mu$ for some $\mu \in (0, 1)$. In addition,

$$|\mathbb{E}_k[v_{i,k}] - v_{i,k}| \leq \kappa_v(\mathbb{E}_k\left[g_{i,k}^2\right] + g_{i,k}^2) \tag{3.1}$$

for some $\kappa_v > 0$ and all $k \geq 0$.

Assumption 3.2. For every realization of the algorithm, we have that $g_{i,k}^2 \leq v_{i,k}$ for all $i \in \{1, \ldots, n\}$ and all $k \geq 0$.

We immediately note that Assumption 3.1 implies that

$$v_{i,k} \geq \min_{i \in \{1,\ldots,n\}} \varsigma_i \stackrel{\text{def}}{=} \varsigma_{\min} \tag{3.2}$$

and Assumption 3.2 ensures that

$$\mathbb{E}_k\left[g_{i,k}^2\right] \leq \mathbb{E}_k[v_{i,k}]\,. \tag{3.3}$$

The first step in our analysis is to derive a parametric bound on the decrease in the exact linear model of F caused by the step s_k, using a technique inspired by [41] and [16].

[(1)] Not to mention that an infinite gradient is likely to crash the algorithm on many machines.

Lemma 3.3. *Let s_j^L be the step produced at the j-th iteration by the* ASGRAD *algorithm. Suppose also that Assumptions 2.4, 3.1 and 3.2 hold. Let G_j be the true gradient of F at x_j. Then, for all $i \in \{1, \ldots, n\}$,*

$$\mathbb{E}_j\left[\gamma_j G_{i,j} s_{i,j}^L\right] \leq -(1-\frac{\mu}{2})\frac{\gamma_{\text{low}} G_{i,j}^2}{(\mathbb{E}_j[v_{i,j}])^\mu} + 2\kappa_\Delta \mathbb{E}_j\left[\frac{g_{i,j}^2}{w_{i,j}^2}\right], \tag{3.4}$$

where

$$\kappa_\Delta \stackrel{\text{def}}{=} \frac{\mu\kappa_v^2}{\gamma_{\text{low}}}\left[\kappa_g^{2\mu} + \frac{\kappa_g^2}{\varsigma_{\min}^{1-\mu}} + \frac{\kappa_g^{4-2\mu}}{\varsigma_{\min}^{2-2\mu}} + \kappa_g^{2-2\mu}\kappa_\mu\right] \quad \textit{with} \quad \kappa_\mu \stackrel{\text{def}}{=} \frac{1}{\varsigma_{\min}^{1-2\mu}}\mathbb{1}_{\mu<\frac{1}{2}} + \kappa_g^{4\mu-2}\mathbb{1}_{\mu\geq\frac{1}{2}}, \tag{3.5}$$

where $\mathbb{1}_{\mathcal{E}}$ stands for the indicator function of the event $\mathcal{E}$.

Proof. See Appendix.

This lemma essentially implies that s^L provides a descent direction on the true F as long as the square of the true gradient's norm remains large compared with the stepsizes. We also need another result, partly inspired by [16, 41], whose utility will be to bound the last term on the right-hand side of (3.4).

Lemma 3.4. *Let $\{a_k\}_{k\geq 0}$ be a non-negative sequence, $\alpha > 0$ and define, for each $k \geq 0$, $b_k = \sum_{j=0}^k a_j$. Then if $\alpha \neq 1$,*

$$\sum_{j=0}^k \frac{a_j}{(\varsigma + b_j)^\alpha} \leq \frac{1}{(1-\alpha)}((\varsigma + b_k)^{1-\alpha} - \varsigma^{1-\alpha}). \tag{3.6}$$

Otherwise (i.e. if $\alpha = 1$) (see Lemma 5.2 in [16]),

$$\sum_{j=0}^k \frac{a_j}{\varsigma + b_j} \leq \log\left(\frac{\varsigma + b_k}{\varsigma}\right). \tag{3.7}$$

Proof. See Appendix. Note that (3.7) is the limit of (3.6) when α tends to one.

Using both Lemmas 3.3 and 3.4, we are now in position to deduce a first result on the global convergence rate of a class of ASGRAD algorithms using specific "Adagrad-like" scaling factors satisfying Assumptions 3.1 and 3.2.

Theorem 3.5. *Suppose that Assumptions 2.1–2.4 hold and that the* ASGRAD *algorithm is applied to problem (2.1) where, for all $k \geq 0$ and all $i \in \{1, \ldots, n\}$,*

$$w_{i,k} = \left(\varsigma + \sum_{\ell=0}^k g_{i,\ell}^2\right)^\mu, \tag{3.8}$$

where $\varsigma \in (0, \kappa_g]$ and $\mu \in (0, 1)$. Then the following bounds hold for κ_Δ given in (3.5) and

$$\kappa_\square \stackrel{\text{def}}{=} \frac{\kappa_g^{2\mu}(4\kappa_\Delta + L)}{(1-\frac{\mu}{2})\gamma_{\text{low}}}. \tag{3.9}$$

(i) If $\mu \in (0, \frac{1}{2})$, then

$$\mathbb{E}\left[\operatorname*{average}_{j\in\{0,\dots,k\}} \|G_j\|^2\right] \leq \frac{2\kappa_g^{2\mu}}{(1-\frac{\mu}{2})\gamma_{\text{low}}(k+1)^{1-\mu}}\Big[F(x_0) - F_{\text{low}}\Big] + \frac{n\kappa_\square}{1-2\mu}\frac{(\varsigma+\kappa_g^2(k+1))^{1-2\mu} - \varsigma^{1-2\mu}}{(k+1)^{1-\mu}}. \tag{3.10}$$

(ii) If $\mu = \frac{1}{2}$, then

$$\mathbb{E}\left[\operatorname*{average}_{j\in\{0,\dots,k\}} \|G_j\|^2\right] \leq \frac{8\kappa_g}{3\gamma_{\text{low}}\sqrt{(k+1)}}\Big[F(x_0) - F_{\text{low}}\Big] + n\kappa_\square\frac{\log\left(1+(k+1)\frac{\kappa_g^2}{\varsigma}\right)}{\sqrt{(k+1)}}. \tag{3.11}$$

(iii) If $\mu \in (\frac{1}{2}, 1)$, then

$$\mathbb{E}\left[\operatorname*{average}_{j\in\{0,\dots,k\}} \|G_j\|^2\right] \leq \frac{2\kappa_g^{2\mu}}{(1-\frac{\mu}{2})\gamma_{\text{low}}(k+1)^{1-\mu}}\Big[F(x_0) - F_{\text{low}}\Big] + \frac{n\kappa_\square}{2\mu-1}\frac{\varsigma^{1-2\mu} - (\varsigma+\kappa_g^2(k+1))^{1-2\mu}}{(k+1)^{1-\mu}}. \tag{3.12}$$

Proof. It is clear from (3.8) that $w_{i,k} \geq \varsigma^\mu$. Moreover, if we define $v_{i,k} \overset{\text{def}}{=} \varsigma + \sum_{\ell=0}^k g_{i,\ell}^2$, then we have that $w_{i,k} = v_{i,k}^\mu$, $v_{i,k} \geq g_{i,k}^2$ and

$$|\mathbb{E}_k[v_{i,k}] - v_{i,k}| = |\mathbb{E}_k\big[g_{i,k}^2\big] - g_{i,k}^2| \leq \mathbb{E}_k\big[g_{i,k}^2\big] + g_{i,k}^2.$$

Thus the proposed scaling factors verify Assumptions 3.1 and 3.2 with $\kappa_v = 1$.
Using (2.2), we derive that

$$F(x_{j+1}) \leq F(x_j) + \gamma_j G_j^T s_j^L + \frac{L}{2}\gamma_j^2\|s_j^L\|^2 \leq F(x_j) + \gamma_j G_j^T s_j^L + \frac{L}{2}\|s_j^L\|^2.$$

Taking the conditional expectation, using Lemma 3.3, the fact that $v_{i,j} \leq (k+2)\kappa_g^2$ (because we assumed that $\varsigma \leq \kappa_g$), (2.4), we deduce that, for $j \in \{0, \dots, k\}$,

$$\begin{aligned}
\mathbb{E}_j[F(x_{j+1})] &\leq F(x_j) + \sum_{i=1}^n \mathbb{E}_j\big[\gamma_j G_{i,j} s_{i,j}^L\big] + \frac{L}{2}\mathbb{E}_j\big[\|s_j^L\|^2\big], \\
&\leq F(x_j) - \sum_{i=1}^n (1-\frac{\mu}{2})\gamma_{\text{low}}\frac{G_{i,j}^2}{(\mathbb{E}_j[v_{i,j}])^\mu} + 2\kappa_\Delta\mathbb{E}_j\left[\frac{g_{i,j}^2}{w_{i,j}^2}\right] + \frac{L}{2}\mathbb{E}_j\big[\|s_j^L\|^2\big], \\
&\leq F(x_j) - (1-\frac{\mu}{2})\gamma_{\text{low}}\frac{\|G_j\|^2}{\kappa_g^{2\mu}(k+2)^\mu} + \left(\frac{L}{2} + 2\kappa_\Delta\right)\mathbb{E}_j\big[\|s_j^L\|^2\big].
\end{aligned}$$

We may now take the full expectation and sum the previous inequality for $j \in \{0, \dots, k\}$ to derive that

$$\begin{aligned}\mathbb{E}[F(x_{k+1})] &\leq F(x_0) - (1-\frac{\mu}{2})\frac{\gamma_{\text{low}}}{\kappa_g^{2\mu}(k+2)^\mu}\sum_{j=0}^{k}\mathbb{E}\big[\|G_j\|^2\big] + \left(\frac{L}{2}+2\kappa_\Delta\right)\sum_{j=0}^{k}\mathbb{E}\big[\|s_j^L\|^2\big] \\ &\leq F(x_0) - (1-\frac{\mu}{2})\frac{\gamma_{\text{low}}}{\kappa_g^{2\mu}(k+2)^\mu}\sum_{j=0}^{k}\mathbb{E}\big[\|G_j\|^2\big] + \left(\frac{L}{2}+2\kappa_\Delta\right)\sum_{i=1}^{n}\sum_{j=0}^{k}\mathbb{E}\big[(s_{i,j}^L)^2\big]\,.\end{aligned} \tag{3.13}$$

Using now Lemma 3.4 with $\alpha = 2\mu$ for each $s_{i,j}^L$, (2.4), (3.8) and Assumption 2.4, we derive that, for $\mu \in (0, \frac{1}{2})$,

$$\begin{aligned}\sum_{j=0}^{k}(s_{i,j}^L)^2 &= \sum_{j=0}^{k}\frac{g_{i,j}^2}{(\varsigma+\sum_{j=0}^{k}g_{i,j}^2)^{2\mu}} \\ &\leq \frac{1}{1-2\mu}\left[\left(\varsigma+\sum_{j=0}^{k}g_{i,j}^2\right)^{1-2\mu} - \varsigma^{1-2\mu}\right] \\ &\leq \frac{1}{1-2\mu}\left[\Big(\varsigma+(k+1)\kappa_g^2\Big)^{1-2\mu} - \varsigma^{1-2\mu}\right].\end{aligned}$$

Plugging this inequality in (3.13) and using Assumption 2.3, we obtain that

$$\begin{aligned}F_{\text{low}} \leq \mathbb{E}[F(x_{k+1})] \leq F(x_0) &- (1-\frac{\mu}{2})\frac{\gamma_{\text{low}}}{\kappa_g^{2\mu}(k+2)^\mu}\sum_{j=0}^{k}\mathbb{E}\big[\|G_j\|^2\big] \\ &+ \frac{n}{1-2\mu}\left(\frac{L}{2}+2\kappa_\Delta\right)\left[(\varsigma+(k+1)\kappa_g^2)^{1-2\mu} - \varsigma^{1-2\mu}\right]\end{aligned}$$

and thus, since $(k+2)^\mu \leq 2(k+1)^\mu$, that

$$(k+1)\mathbb{E}\left[\underset{j\in\{0,\dots,k\}}{\text{average}}\|G_j\|^2\right] \leq \sum_{j=0}^{k}\mathbb{E}\big[\|G_j\|^2\big] \tag{3.14}$$

$$\leq \frac{2\kappa_g^{2\mu}(F(x_0)-F_{\text{low}})}{(1-\frac{\mu}{2})\gamma_{\text{low}}(k+1)^{-\mu}} \tag{3.15}$$

$$+ \frac{n\left[(\varsigma+\kappa_g^2(k+1))^{1-2\mu} - \varsigma^{1-2\mu}\right]}{(1-2\mu)(k+1)^{-\mu}}\left(\frac{\kappa_g^{2\mu}\Big(L+4\kappa_\Delta\Big)}{\gamma_{\text{low}}(1-\frac{\mu}{2})}\right),$$

which is (3.10).
If $\mu = \frac{1}{2}$, we reuse (3.13) and Lemma 3.4 for each $s_{i,j}^L$ with $\alpha = 1$, and derive that, in this case,

$$\mathbb{E}[F(x_{k+1})] \leq F(x_0) - \frac{3}{4}\frac{\gamma_{\text{low}}}{\sqrt{(k+2)}\kappa_g}\sum_{j=0}^{k}\mathbb{E}\big[\|G_j\|^2\big] + n\left(\frac{L}{2}+2\kappa_\Delta\right)\log\left(1+(k+1)\frac{\kappa_g^2}{\varsigma}\right).$$

By a reasoning similar to that leading to (3.14) we now obtain that

$$(k+1)\mathbb{E}\left[\operatorname*{average}_{j\in\{0,\ldots,k\}}\|G_j\|^2\right] \leq \sum_{j=0}^{k}\mathbb{E}\big[\|G_j\|^2\big]$$
$$\leq \left(\frac{4}{3}\right)\frac{2\kappa_g(F(x_0)-F_{\rm low})\sqrt{(k+1)}}{\gamma_{\rm low}}$$
$$+\left(\frac{4n}{3}\right)\frac{\kappa_g}{\gamma_{\rm low}}(L+4\kappa_\Delta)\log\left(1+(k+1)\frac{\kappa_g^2}{\varsigma}\right)\sqrt{(k+1)}.$$

Rearranging the terms yields (3.11).
Finally, if $\mu\in(\frac{1}{2},1)$, we again reuse (3.13) and Lemma 3.4 for each $s_{i,j}^L$ with $\alpha=2\mu>1$, and deduce that

$$\mathbb{E}[F(x_{k+1})] \leq F(x_0)-(1-\frac{\mu}{2})\frac{\gamma_{\rm low}}{(k+2)^\mu\kappa_g^{2\mu}}\sum_{j=0}^{k}\mathbb{E}\big[\|G_j\|^2\big]$$
$$+\left(\frac{L}{2}+2\kappa_\Delta\right)\frac{n}{2\mu-1}\left(\varsigma^{1-2\mu}-(\varsigma+\kappa_g^2(k+1))^{1-2\mu}\right).$$

Following the same argument as above yields that

$$(k+1)\mathbb{E}\left[\operatorname*{average}_{j\in\{0,\ldots,k\}}\|G_j\|^2\right] \leq \sum_{j=0}^{k}\mathbb{E}\big[\|G_j\|^2\big]$$
$$\leq \frac{2\kappa_g^{2\mu}(F(x_0)-F_{\rm low})}{(1-\frac{\mu}{2})\gamma_{\rm low}(k+1)^{-\mu}}+\frac{n}{2\mu-1}\left(\frac{\kappa_g^{2\mu}\Big(L+4\kappa_\Delta\Big)}{\gamma_{\rm low}(1-\frac{\mu}{2})}\right)\times$$
$$\frac{\varsigma^{1-2\mu}-(\varsigma+\kappa_g^2(k+1))^{1-2\mu}}{(k+1)^{-\mu}}.$$

Rearranging the terms gives (3.12). □

Note that the last fractions in the last terms of (3.10) and (3.12) have been written in a form stressing the continuity with (3.11), but could obviously be bounded above by the simpler

$$\frac{(\varsigma+\kappa_g^2)^{1-2\mu}}{(k+1)^\mu} \quad\text{and}\quad \frac{\varsigma^{1-2\mu}}{(k+1)^{1-\mu}}$$

respectively.

Theorem 3.5 suggests a few comments. The first is that (3.10), (3.11) and (3.12) guarantee the convergence of the ASGRAD algorithm with (3.8) to first-order critical points, because their right-hand sides all tend to zero when k tends to infinity. The rate at which this convergence occurs, however, differs for the three cases, depending on the parameter μ. If constants are lumped into a generic $\mathcal{O}(\cdot)$ notation and using that

$$\frac{1}{(k+1)}\mathbb{E}\left[\sum_{j=0}^{k}\|G_j\|\right] \leq \frac{1}{\sqrt{k+1}}\mathbb{E}\left[\sqrt{\sum_{j=0}^{k}\|G_j\|^2}\right] \leq \sqrt{\mathbb{E}\left[\operatorname*{average}_{j\in\{0,\ldots,k\}}\|G_j\|^2\right]}$$

where we used Cauchy Schwartz and Jensen's inequality, we obtain, that

$$\mathbb{E}\left[\underset{j\in\{0,\dots,k\}}{\text{average}}\ \|G_j\|\right] \leq \begin{cases} \mathcal{O}\left(\frac{1}{(k+1)^{\frac{1}{2}\mu}}\right) & (\mu \in (0, \tfrac{1}{2})), \\ \mathcal{O}\left(\frac{\log(k+1)}{(k+1)^{\frac{1}{4}}}\right) & (\mu = \tfrac{1}{2}), \\ \mathcal{O}\left(\frac{1}{(k+1)^{\frac{1}{2}(1-\mu)}}\right) & (\mu \in (\tfrac{1}{2}, 1)). \end{cases}$$

Examining these "k-order" bounds indicates that the best bound is that corresponding to $\mu = \frac{1}{2}$. This is nothing but the standard Adagrad algorithm.

3.1 Comparison with prior work for $\mu = \frac{1}{2}$

To provide more context for the reader and to better locate our work within the vast literature dealing with the theoretical analysis of Adagrad, we now discuss our result for $\mu = \frac{1}{2}$. We immediately note that our bound in $\mathcal{O}\left(\frac{\log(k+1)}{\sqrt{(k+1)}}\right)$ on $\mathbb{E}\Big[\text{average}_{j\in\{0,\dots,k\}}\ \|G_j\|\Big]$ is not new for Adagrad and has been obtained under various assumptions, be it under the requirement of gradient boundedness [16, 28, 41, 45] such as our case, in the unbounded case [19, 40, 20, 1, 31], under various Lipschitz smoothness assumptions, see [20, 32, 40] and broader noise assumptions [1, 27, 28, 31]. These works either focused on a component-wise Adagrad like the one we analyzed, or study a variant called Adagrad-Norm that set a global scaling $w_k \propto \sqrt{\sum_{i=1}^{k} \|g_k\|^2}$ for each dimension. The final result was also presented in high-probability [1, 28, 31] or in expectation just as our case [16, 19, 20, 41]. For a compact summary of the last results for the study of Adagrad variants, see [27, Table 1]. Our contribution is to show that the choice $\mu = \frac{1}{2}$ is optimal within a wide range class of adaptive gradient methods (3.8).

We also note that $\mathcal{O}\left(\frac{\log(k+1)}{\sqrt{(k+1)}}\right)$ differs by a logarithmic term from the convergence rate of well-tuned stochastic first order methods proven in [22], and so improving our current bound on $\mathbb{E}\Big[\text{average}_{j\in\{0,\dots,k\}}\ \|G_j\|\Big]$ may be possible under additional assumptions, such as a tighter variance bound on $g_{i,k}$.

4 A "divergent series" class of ASGRAD algorithms

One might then wonder if a class of ASGRAD algorithms exists where improved asymptotic convergence rate can be achieved. This section considers two cases of interest, both depending on some constants $\mu \in (0, 1)$ and $\varsigma > 0$. The first, which we call *maxgi*, is defined, for some $\alpha > 1$ and $i \in \{1, \dots, n\}$, by

$$w_{i,k} = \xi_{i,k}(k+1)^{\mu} \ \text{ where } \ \xi_{i,k} = \begin{cases} \varsigma & \text{if } \ k = -1, \\ |g_{i,k}| & \text{if } \ k \geq 0 \ \text{ and } \ |g_{i,k}| \geq \alpha\xi_{i,k-1} \\ \xi_{i,k-1} & \text{if } \ k \geq 0 \ \text{ and } \ |g_{i,k}| < \alpha\xi_{i,k-1} \end{cases} \tag{4.1}$$

The second, called *avrgi*, uses for $i \in \{1, \ldots, n\}$

$$w_{i,k} = \xi_{i,k}(k+1)^{\mu} \quad \text{where} \quad \xi_{i,k} = \max\left[\varsigma, \frac{1}{k+1}\sum_{j=0}^{k}|g_{i,j}|\right]. \tag{4.2}$$

Before delving into the analysis, we briefly mention that only $\xi_{i,k}$ defined in (4.1) is a monotonically increasing sequence whereas this is not necessarily the case for (4.2). In both cases, $\xi_{i,k}$ lies in the interval $[\varsigma, \kappa_g]$ if Assumption 2.4 holds
We first state a crucial decrease result.

Lemma 4.1. *Suppose that Assumptions 2.1–2.4 hold and that the* ASGRAD *algorithm is applied to problem (2.1) with its scaling factors being defined, for some $\mu \in (0,1)$ by (4.1) or (4.2). Then*

$$\mathbb{E}_j\left[-\gamma_j \frac{G_{i,j}g_{i,j}}{w_{i,j}}\right] \leq -\kappa_1 \frac{G_{i,j}^2}{(j+1)^{\mu}} + \kappa_2 \frac{\mathbb{P}_j[\mathcal{A}_j]}{(j+1)^{\mu}} + \kappa_3 \mathbb{E}_j\left[\frac{g_{i,j}^2}{w_{i,j}^2}\right], \tag{4.3}$$

where $\mathcal{A}_j$ denotes the event $\{|g_{i,j}| \geq \alpha\xi_{i,j-1}\}$ and

1. *if (4.1) is used,*

$$\kappa_1 = \frac{\gamma_{\text{low}}}{\varsigma}, \quad \kappa_2 = \frac{4\kappa_g^3}{\varsigma^2} \quad \textit{and} \quad \kappa_3 = 0, \tag{4.4}$$

2. *if (4.2) is used,*

$$\kappa_1 = \frac{\gamma_{\text{low}}}{2\varsigma}, \quad \kappa_2 = 0 \quad \textit{and} \quad \kappa_3 = \frac{2\kappa_g^2}{\varsigma\gamma_{\text{low}}}. \tag{4.5}$$

Proof. The proof is somewhat technical and given in Appendix.

Observe that, although it is well defined for both cases, the event $\mathcal{A}_j$ is only relevant for *maxgi* because $\kappa_2 = 0$ in (4.5). We also give an important property for the *maxgi* case.

Lemma 4.2. *Suppose that Assumptions 2.1 and 2.4 hold and that the* ASGRAD *algorithm is applied to problem (2.1) with its scaling factors being defined by (4.1) for some $\mu \in (0,1)$. Then, for all $k \geq 0$,*

$$\mathbb{E}\left[\sum_{j=0}^{k}\mathbb{P}_j[\mathcal{A}_j]\right] \leq \left\lfloor \frac{\log(\kappa_g) - \log(\varsigma)}{\log(\alpha)} \right\rfloor \stackrel{\text{def}}{=} \tau_{\max}. \tag{4.6}$$

Proof. For $k \geq 0$, let τ_k be the number of occurrences of $\mathcal{A}_j$ for $j \in \{0, \ldots, k\}$. We must have that, for all j,

$$\kappa_g \geq \xi_{i,j} \geq \xi_{i,-1}\alpha^{\tau_k} = \varsigma\alpha^{\tau_k}$$

and hence, for all k,

$$\sum_{j=0}^{k}\mathbb{1}_{\mathcal{A}_j} = \tau_k \leq \left\lfloor \frac{\log(\kappa_g) - \log(\varsigma)}{\log(\alpha)} \right\rfloor. \tag{4.7}$$

Using this bound and the law of total expectation, we thus obtain that for all k,

$$\mathbb{E}\left[\sum_{j=0}^{k}\mathbb{P}_j[\mathcal{A}_j]\right] = \sum_{j=0}^{k}\mathbb{E}\big[\mathbb{E}_j\big[\mathbb{1}_{\mathcal{A}_j}\big]\big] = \sum_{j=0}^{k}\mathbb{E}\big[\mathbb{1}_{\mathcal{A}_j}\big] = \mathbb{E}\left[\sum_{j=0}^{k}\mathbb{1}_{\mathcal{A}_j}\right] \leq \left\lfloor \frac{\log(\kappa_g) - \log(\varsigma)}{\log(\alpha)} \right\rfloor.$$

□

Theorem 4.3. *Suppose that Assumptions 2.1–2.4 hold and that the* ASGRAD *algorithm is applied to problem (2.1) with its scaling factors being defined, for some $\mu \in (0,1)$ by (4.1) or (4.2). Then, for $\mu \neq \frac{1}{2}$,*

$$\begin{aligned}\mathbb{E}\Big[\text{average}_{j\in\{0,\ldots,k\}} \|G_j\|^2\Big] &\leq \frac{F(x_0)+n\kappa_2\tau_{\max}-F_{\low}}{\kappa_1(k+1)^{1-\mu}} \\ &\quad + \frac{n\kappa_g^2}{\kappa_1\varsigma^2(1-2\mu)}\left[\kappa_3+\frac{L}{2}\right]\left[\frac{1}{(k+1)^{\mu}}-\frac{2\mu}{(k+1)^{1-\mu}}\right],\end{aligned} \tag{4.8}$$

while, if $\mu = \frac{1}{2}$,

$$\mathbb{E}\left[\underset{j\in\{0,\ldots,k\}}{\text{average}} \|G_j\|^2\right] \leq \frac{F(x_0)+n\kappa_2\tau_{\max}-F_{\low}}{\kappa_1\sqrt{k+1}} + \frac{n\kappa_g^2}{\kappa_1\varsigma^2}\left[\kappa_3+\frac{L}{2}\right]\frac{1+\log(k+1)}{\sqrt{k+1}}, \tag{4.9}$$

where κ_1, κ_2, κ_3 and $\tau_{\max}$ are defined in Lemmas 4.1 and 4.2.

Proof. By using (2.2), the inequality $\gamma_j \leq 1$, and (2.4), we derive that

$$F(x_{j+1}) \leq F(x_j) + \gamma_j G_j^T s_j^L + \frac{L}{2}\gamma_j^2\|s_j^L\|^2 \leq F(x_j) - \gamma_j\frac{G_j^T g_{i,j}}{w_{i,j}} + \frac{L}{2}\sum_{i=1}^n \frac{g_{i,j}^2}{w_{i,j}^2}. \tag{4.10}$$

Using Lemma 4.1, taking the conditional expectation of (4.10) and using Assumption 2.4 to bound $g_{i,j}^2$ and that $w_{i,k} \geq \varsigma(k+1)^{\mu}$ for both choices (4.1) and (4.2), we obtain that

$$\begin{aligned}\mathbb{E}_j[F(x_{j+1})] &\leq F(x_j) + \sum_{i=1}^n \mathbb{E}_j\left[\gamma_j G_{i,j}\frac{g_{i,j}}{w_{i,j}}\right] + \frac{L}{2}\mathbb{E}_j\left[\frac{g_{i,j}^2}{w_{i,j}^2}\right], \\ &\leq F(x_j) - \sum_{i=1}^n \left[\kappa_1\frac{G_{i,j}^2}{(j+1)^{\mu}} + \kappa_2\frac{\mathbb{P}_j[\mathcal{A}_j]}{(j+1)^{\mu}} + \kappa_3\mathbb{E}_j\left[\frac{g_{i,k}^2}{w_{i,j}^2}\right] + \frac{L}{2}\mathbb{E}_j\left[\frac{g_{i,j}^2}{w_{i,j}^2}\right]\right] \\ &\leq F(x_j) - \sum_{i=1}^n \frac{\kappa_1 G_{i,j}^2}{(j+1)^{\mu}} + n\kappa_2\mathbb{P}_j[\mathcal{A}_j] + \frac{n\kappa_g^2}{\varsigma^2}\left[\kappa_3+\frac{L}{2}\right]\frac{1}{(j+1)^{2\mu}}\end{aligned} \tag{4.11}$$

Summing over all iterations from 0 to k, taking the full expectation and using Lemma 4.2 gives that

$$\begin{aligned}\mathbb{E}[F(x_{k+1})] &\leq F(x_0) - \kappa_1\sum_{j=0}^k\sum_{i=1}^n \frac{\mathbb{E}\big[G_{i,j}^2\big]}{(j+1)^{\mu}} + n\kappa_2\mathbb{E}\left[\sum_{j=0}^k \mathbb{P}_j[\mathcal{A}_j]\right] + \frac{n\kappa_g^2}{\varsigma^2}\left[\kappa_3+\frac{L}{2}\right]\sum_{j=0}^k\frac{1}{(j+1)^{2\mu}} \\ &\leq F(x_0) + n\kappa_2\tau_{\max} - \kappa_1\sum_{j=0}^k\sum_{i=1}^n \frac{\mathbb{E}\big[G_{i,j}^2\big]}{(j+1)^{\mu}} + \frac{n\kappa_g^2}{\varsigma^2}\left[\kappa_3+\frac{L}{2}\right]\sum_{j=0}^k\frac{1}{(j+1)^{2\mu}}\end{aligned}$$

If we now define

$$\phi_{\mu}(x) \stackrel{\text{def}}{=} \begin{cases} \dfrac{(x+1)^{1-2\mu}-1}{1-2\mu} & \text{if} \quad \mu \neq \frac{1}{2} \\ \log(x+1) & \text{otherwise,}\end{cases}$$

we may bound the last inequality, using a simple sum-integral comparison and Assumption 2.3 to obtain that

$$\sum_{j=0}^k\sum_{i=1}^n \mathbb{E}\big[G_{i,j}^2\big] \leq \frac{(k+1)^{\mu}(F(x_0)+n\kappa_2\tau_{\max}-F_{\low})}{\kappa_1} + \frac{n\kappa_g^2}{\varsigma^2}\left[\kappa_3+\frac{L}{2}\right]\frac{(k+1)^{\mu}}{\kappa_1}(1+\phi_{\mu}(k))$$

and thus that

$$\mathbb{E}\left[\underset{j\in\{0,\ldots,k\}}{\text{average}}\ \|G_j\|^2\right] \leq \frac{F(x_0)+n\kappa_2\tau_{\max}-F_{\low}}{\kappa_1(k+1)^{1-\mu}} + \frac{n\kappa_g^2}{\kappa_1\varsigma^2}\left[\kappa_3+\frac{L}{2}\right]\frac{1+\phi_\mu(k)}{(k+1)^{1-\mu}}.$$

This gives (4.9) when $\mu=\frac{1}{2}$. Otherwise, (4.8) follows from the fact that

$$1+\phi_\mu(k) = \frac{1}{1-2\mu}\left[\frac{1}{(k+1)^{2\mu-1}}-2\mu\right].$$

□

The choices (4.1) and (4.2) are of course reminiscent, in a smooth but stochastic and nonconvex setting, of the "divergent stepsize" subgradient method for non-smooth convex optimization (see [3] and the many references therein), for which a $\mathcal{O}(1/\sqrt{k})$ global rate of convergence is known (Theorems 8.13 and 8.30 in this last reference).

The bounds given by Theorem 4.3 are qualitatively similar to those of Theorem 3.5, but they may be improved if we strengthen our assumptions, and impose an additional conditional variance condition on the gradient estimator.

Theorem 4.4. *Suppose that Assumptions 2.1–2.4 hold and that an* ASGRAD *algorithm is applied to problem (2.1) with its scaling factors being defined (4.1) and (4.2). Suppose also that, for all $i\in\{1,\ldots,n\}$ and all $k\geq 0$*

$$\mathrm{Var}_k\left[g_{i,k}\right] = \mathbb{E}_k\left[g_{i,k}^2-G_{i,k}^2\right] \leq \kappa_{\var}G_{i,k}^2 \tag{4.12}$$

holds for some $\kappa_{\var}\geq 0$. Then, for any $\theta\in(0,\kappa_1)$,

$$\mathbb{E}\left[\underset{j\in\{j_\theta+1,\ldots,k\}}{\text{average}}\ \|G_j\|^2\right] \leq \kappa_\#(\theta)\frac{(k+1)^\mu}{k-j_\theta} \leq \frac{\kappa_\#(\theta)(j_\theta+2)}{(k+1)^{1-\mu}}, \tag{4.13}$$

where

$$\kappa_\#(\theta) \overset{\text{def}}{=} \frac{1}{\theta}\left(F(x_0)+n\kappa_2\tau_{\max}-F_{\low}+\frac{n\,2^\mu\kappa_g^4}{\varsigma^4}\left[\kappa_3+\frac{L}{2}\right](1+\kappa_{\var})j_\theta\right), \tag{4.14}$$

and

$$j_\theta \overset{\text{def}}{=} \left\lfloor\left(\left[\kappa_3+\frac{L}{2}\right]\frac{\kappa_g^2 2^\mu}{\varsigma^4(\kappa_1-\theta)}\right)^{\frac{1}{\mu}}\right\rfloor + 1. \tag{4.15}$$

Proof. To simplify notation, set, for the course of this proof, $w_{i,-1}=\varsigma$, $i\in\{1,\ldots,n\}$, $\frac{0}{0}=1$. As in the proof of Theorem 4.3, we derive (see (4.11)) that

$$\begin{aligned}
\mathbb{E}_j[F(x_{j+1})] &\leq F(x_j)-\sum_{i=1}^n\left[\kappa_1\frac{G_{i,j}^2}{(j+1)^\mu}+\kappa_2\frac{\mathbb{P}_j[\mathcal{A}_j]}{(j+1)^\mu}+\left[\kappa_3+\frac{L}{2}\right]\mathbb{E}_j\left[\frac{g_{i,j}^2}{w_{i,j}^2}\right]\right]\\
&\leq F(x_j)-\sum_{i=1}^n\left[\frac{\kappa_1 G_{i,j}^2}{(j+1)^\mu}+\kappa_2\frac{\mathbb{P}_j[\mathcal{A}_j]}{(j+1)^\mu}+\left[\kappa_3+\frac{L}{2}\right]\mathbb{E}_j\left[\frac{g_{i,j}^2}{w_{i,j-1}^2}\left(\frac{w_{i,j-1}}{w_{i,j}}\right)^2\right]\right]\\
&\leq F(x_j)-\sum_{i=1}^n\left[\frac{\kappa_1 G_{i,j}^2}{(j+1)^\mu}+\kappa_2\frac{\mathbb{P}_j[\mathcal{A}_j]}{(j+1)^\mu}+\left[\kappa_3+\frac{L}{2}\right]\frac{\kappa_g^2}{\varsigma^2}\mathbb{E}_j\left[\frac{g_{i,j}^2}{w_{i,j-1}^2}\right]\right]\\
&\leq F(x_j)-\sum_{i=1}^n\left[\frac{\kappa_1 G_{i,j}^2}{(j+1)^\mu}+\kappa_2\mathbb{P}_j[\mathcal{A}_j]+\left[\kappa_3+\frac{L}{2}\right]\frac{\kappa_g^2(1+\kappa_{\var})}{\varsigma^2 w_{i,j-1}^2}G_{i,j}^2\right],
\end{aligned}$$

where we have used the fact that $\left(\frac{w_{i,j-1}}{w_{i,j}}\right)^2 \leq \frac{\kappa_g^2}{\varsigma^2}$ (because of (4.1) and (4.2)), the measurability of $w_{i,j-1}$ with respect to the past and (4.12) to deduce the last inequality. Using now the bound $\frac{(j+1)^\mu}{w_{i,j-1}} \leq \frac{2^\mu}{\varsigma}$ and summing over the iterations for $j \in \{0, \ldots, k\}$ then yields that

$$\sum_{j=0}^{k} \mathbb{E}_j[F(x_{j+1})] \leq \sum_{j=0}^{k} F(x_j) + n\kappa_2 \sum_{j=0}^{k} \mathbb{P}_j[\mathcal{A}_j] + \sum_{j=0}^{k} \sum_{i=1}^{n} \frac{G_{i,j}^2}{(j+1)^\mu} \left(-\kappa_1 + \frac{\widehat{\kappa}}{w_{i,j-1}}\right) \tag{4.16}$$

with $\widehat{\kappa} = \left[\kappa_3 + \frac{L}{2}\right] \frac{\kappa_g^2 2^\mu}{\varsigma^3}(1 + \kappa_{\text{var}})$. Note now that the definition of j_θ in (4.15) and the fact that $w_{i,j-1} \geq \varsigma\, j^\mu$ together imply that

$$\left(-\kappa_1 + \frac{\widehat{\kappa}}{w_{i,j-1}}\right) \leq -\theta, \tag{4.17}$$

for $j \geq j_\theta$. Hence, from (4.16),

$$\begin{aligned} \textstyle\sum_{j=0}^{k} \mathbb{E}_j[F(x_{j+1})] \leq \; & \sum_{j=0}^{k} F(x_j) + n\kappa_2 \sum_{j=0}^{k} \mathbb{P}_j[\mathcal{A}_j] - \theta \sum_{j=j_\theta}^{k} \sum_{i=1}^{n} \frac{G_{i,j}^2}{(j+1)^\mu} \\ & + \sum_{j=0}^{j_\theta - 1} \sum_{i=1}^{n} \frac{G_{i,j}^2}{(j+1)^\mu} \left(-\kappa_1 + \frac{\widehat{\kappa}}{w_{i,j-1}}\right), \end{aligned} \tag{4.18}$$

and the last term of this inequality is bounded by

$$\sum_{j=0}^{j_\theta - 1} \sum_{i=1}^{n} \frac{G_{i,j}^2}{(j+1)^\mu} \left(-\kappa_1 + \frac{\widehat{\kappa}}{w_{i,j}}\right) \leq \sum_{j=0}^{j_\theta - 1} \sum_{i=1}^{n} \widehat{\kappa} \frac{G_{i,j}^2}{\varsigma} \leq \frac{n\kappa_g^2 \widehat{\kappa}}{\varsigma} j_\theta, \tag{4.19}$$

where we used the facts that $\|G\|_\infty \leq \kappa_g$ (because of (2.6)), $w_{i,j} \geq \varsigma$ (because of (4.1) and (4.2)). Injecting (4.19) in (4.18), we deduce that

$$\theta \sum_{j=j_\theta}^{k} \sum_{i=1}^{n} \frac{G_{i,j}^2}{(j+1)^\mu} \leq \sum_{j=0}^{k} F(x_j) + \kappa_2 \sum_{j=0}^{k} \mathbb{P}_j[\mathcal{A}_j] - \sum_{j=0}^{k} \mathbb{E}_j[F(x_{j+1})] + \frac{n\kappa_g^2 \widehat{\kappa}}{\varsigma} j_\theta.$$

Taking the full expectation and using Lemma 4.2 whenever $\kappa_2 > 0$ (i.e. for *maxgi*) then gives that

$$(k - j_\theta)\mathbb{E}\left[\underset{j \in \{j_\theta+1,\ldots,k\}}{\text{average}} \|G_j\|^2\right] \leq \mathbb{E}\left[\sum_{j=j_\theta}^{k} \sum_{i=1}^{n} G_{i,j}^2\right] \leq \frac{(k+1)^\mu}{\theta} \left[F(x_0) + n\kappa_2 \tau_{\max} - F_{\text{low}} + \frac{n\kappa_g^2 \widehat{\kappa}}{\varsigma} j_\theta\right]. \tag{4.20}$$

which gives the desired result. □

The (asymptotic) k-order of convergence of $\mathbb{E}\left[\text{average}_{j \in \{j_\theta+1,\ldots,k\}} \|G_j\|\right]$ implied by (4.13) is therefore

$$\mathcal{O}\left(\frac{1}{(k+1)^{\frac{1}{2}(1-\mu)}}\right)$$

where j_θ is given by (4.15).

5 Numerical illustration

We now provide some numerical illustrations of the algorithmic variants discussed in the previous sections. We trained a simple convolutional network of [23] (denoted in the paper as cifar10-nv) and a small resnet18 model [25] on the CIFAR-10 image classification dataset[(2)]. For these experiments, we used haiku [26] and optax [2], two JAX [11] based libraries, on a workstation with four GTX 1080TI. We now compare the numerical performance of (3.8) for various μ values in $(0.1, 0.5, 0.9)$ and of the two scaling factor defined by (4.1) and (4.2) with $\mu = 0.1$ $\alpha = 1.1$ and $\varsigma = 0.01$. For the experiments, we have chosen two different learning-rate strategies. For the cifar10-nv architecture, we chose a fixed[(3)] learning rate policy with $\gamma_k = \gamma = 5.\{10^{-4}, 10^{-5}\}$ for all $k \geq 0$.For the resnet18 numerical test, we choose a linearly declining learning rate from $\gamma_{\max} = 5.10^{-2}$ to $\gamma_{\text{low}} = 5.10^{-4}$. Note that this choice is covered by our proposed ASGRAD framework and is considered as a good choice of schedule for the learning rate as it is built in Optax. We used the same random initialization for all scaling choices and followed the data-augmentation procedure of [23], both for training and testing. We trained the models for a total of 100000 steps with a batchsize of 128 using the mean-cross entropy loss function. We report the training and test accuracies (the latter on a sample of size 128 from the test dataset) every 500 steps.

The results of these experiments (averaged over three random runs) are presented in Figures 5.1–5.3. In each figure, the top panel shows the evolution (as a function of the number of steps) of the training accuracy, and the bottom panel that of the test accuracy. The average values are shown as thick lines and the shaded areas of corresponding colour give the 67% confidence intervals.

These simple numerical illustrations are obviously not meant to replace significant numerical testing, but, albeit caution must be exercised not to extrapolate from limited data, they still suggest a few tentative comments.

- The relative behaviour of the tested variants differs significantly between the two tested network architectures, even if the test accuracy is (as expected) slighly lower for the resnet18 case.
- For fixed learning rates, the methods *maxgi* and *avrgi* of the second ASGRAD class (introduced in Section 4) seem to produce relatively good results on our example for fixed learning rate and for the civar-nv architecture, both in training and testing, often outperforming the Adagrad-like variants of the first class (of Section 3).
- Among Adagrad-like variants, those with a larger μ handle smaller and fixed learning rates better on these examples, a behaviour admittedly not predicted by our theory.
- The choice of a learning rate schedule for a specific architecture has an impact in practice. We see that for the resnet18 architecture (Figure 5.3), all the methods behave very comparably (except for one variant).
- The comparison of Figures 5.1, 5.2 and Figure 5.3 unsurprisingly shows that, albeit our theory does not depend on the choice of γ_k, the practical convergence behaviour may be affected by this choice (and other factors such as the batchsize).

[(2)] https://www.cs.toronto.edu/ kriz/cifar.html

[(3)] Our choice of a fixed learning rate policy is meant to showcase some intrinsic properties of each scaling factor option.

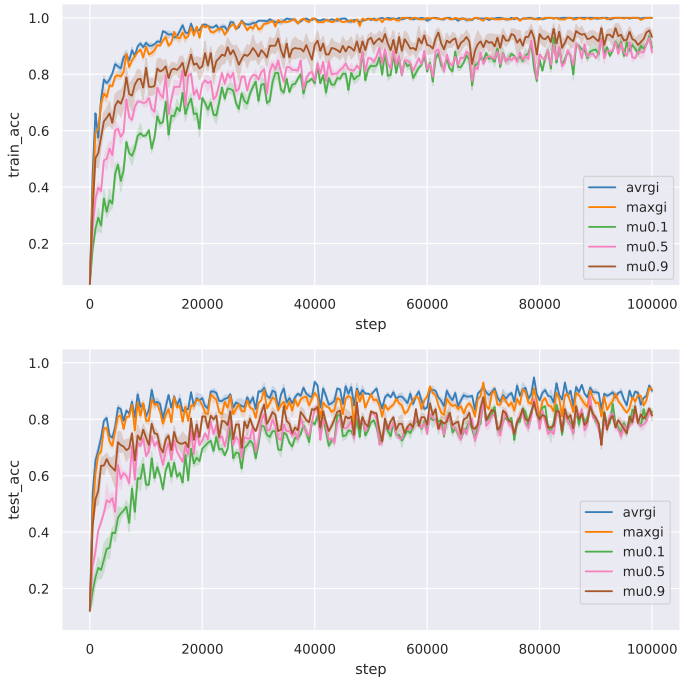

Figure 5.1: Training (top) and test (bottom) accuracies for the Adagrad-like ($\mu \in (0.1, 0.5, 0.9)$), *maxgi* and *avrgi* variants with $\gamma = 5.10^{-4}$ on the cifar10-nv architecture

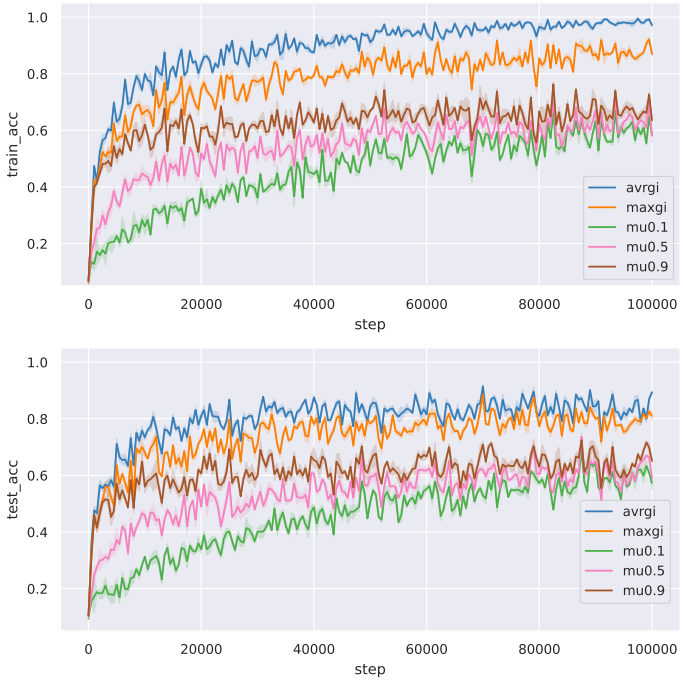

Figure 5.2: Training (top) and test (bottom) accuracy for the Adagrad-like ($\mu \in (0.1, 0.5, 0.9)$), *maxgi* and *avrgi* variants with $\gamma = 5.10^{-5}$ on the cifar10-nv architecture

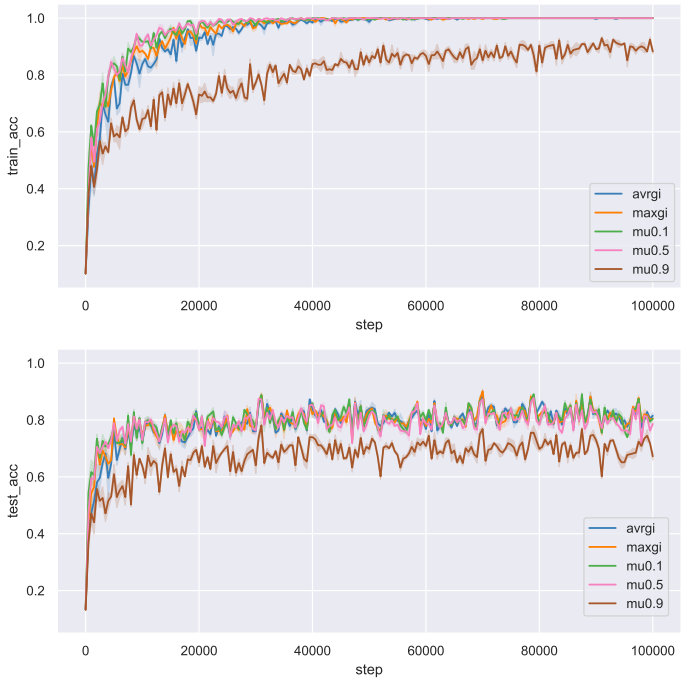

Figure 5.3: Training (top) and test (bottom) accuracies for the Adagrad-like ($\mu \in (0.1, 0.5, 0.9)$), *maxgi* and *avrgi* variants with linearly decaying γ on the resnet18 architecture

6 Conclusions

We have introduced a first-order trust-region framework for minimization methods and derived complexity upper bounds for two classes of interest, the first containing the standard Adagrad. These bounds give the best complexity to values of the class parameters corresponding to Adagrad in the first class. We have also shown these bounds can be improved for both classes under an additional variance condition, in which case the parameter choice yielding the best bounds no longer corresponds to Adagrad. This improvement is asymptotic and implicit for the first class and explicit for the second. However, our numerical illustrations of the discussed methods on examples arising from deep-learning applications indicate that methods of the second class have merits, but also that, at least in our examples, there remains some distance from the above theory to real behaviour. This may possibly be because the complexity bounds may not be sharp, but also, fortunately, because the worst-case happens very rarely in practice.

Acknowledgments This work was supported in part by 3IA Artificial and Natural Intelligence Toulouse Institute, French "Investing for the Future - PIA3" program under the Grant agreement ANR-19-PI3A-0004". The experiments presented in this paper were carried out using the OSIRIM platform that is administered by IRIT and supported by CNRS, the Region Midi-Pyrénées, the French Government, and ERDF (see http://osirim.irit.fr/site/en).

References

[1] Amit Attia and Tomer Koren. SGD with ADAGRAD stepsizes: full adaptivity with high probability to unknown parameters, unbounded gradients and affine variance. In *Proceedings of the 40th International*

Conference on Machine Learning, 2023.

[2] Igor Babuschkin, Kate Baumli, Alison Bell, Surya Bhupatiraju, Jake Bruce, Peter Buchlovsky, David Budden, Trevor Cai, Aidan Clark, Ivo Danihelka, Antoine Dedieu, Claudio Fantacci, Jonathan Godwin, Chris Jones, Ross Hemsley, Tom Hennigan, Matteo Hessel, Shaobo Hou, Steven Kapturowski, Thomas Keck, Iurii Kemaev, Michael King, Markus Kunesch, Lena Martens, Hamza Merzic, Vladimir Mikulik, Tamara Norman, George Papamakarios, John Quan, Roman Ring, Francisco Ruiz, Alvaro Sanchez, Rosalia Schneider, Eren Sezener, Stephen Spencer, Srivatsan Srinivasan, Wojciech Stokowiec, Luyu Wang, Guangyao Zhou, and Fabio Viola. Optax a gradient processing and optimization library for JAX, 2020.

[3] A. Beck. *First-order Methods in Optimization*. Number 25 in MOS-SIAM Optimization Series. SIAM, Philadelphia, USA, 2017.

[4] S. Bellavia, G. Gurioli, B. Morini, and Ph. L. Toint. The impact of noise on evaluation complexity: The deterministic trust-region case. *Journal of Optimization Theory and Applications*, 196(2):700–729, 2023.

[5] Stefania Bellavia, Gianmarco Gurioli, Benedetta Morini, and Philippe L. Toint. A stochastic arc method with inexact function and random derivatives evaluations. In *Thirty-seventh International Conference on Machine Learning: ICML2020*, 2020.

[6] Stefania Bellavia, Gianmarco Gurioli, Benedetta Morini, and Philippe L. Toint. Adaptive regularization for nonconvex optimization using inexact function values and randomly perturbed derivatives. *Journal of Complexity*, 68:91–105, 2022.

[7] Stefania Bellavia, Gianmarco Gurioli, Benedetta Morini, and Philippe L. Toint. Adaptive regularization algorithms with inexact evaluations for nonconvex optimization. *SIAM Journal on Optimization*, 29(4):2881–2915, 2019.

[8] Stefania Bellavia, Gianmarco Gurioli, Benedetta Morini, and Philippe L. Toint. Quadratic and cubic regularisation methods with inexact function and random derivatives for finite-sum minimisation. In *2021 21st International Conference on Computational Science and Its Applications (ICCSA)*. IEEE, September 2021.

[9] Albert S. Berahas, L. Cao, and Katya Scheinberg. Global convergence rate analysis of a generic line search algorithm with noise. *SIAM Journal on Optimization*, 31(2):1489–1518, 2021.

[10] Jose Blanchet, Coralia Cartis, Matt Menickelly, and Katya Scheinberg. Convergence rate analysis of a stochastic trust-region method via supermartingales. *INFORMS Journal on Optimization*, 1(2):92–119, 2019.

[11] James Bradbury, Roy Frostig, Peter Hawkins, Matthew James Johnson, Chris Leary, Dougal Maclaurin, George Necula, Adam Paszke, Jake VanderPlas, Skye Wanderman-Milne, and Qiao Zhang. JAX: composable transformations of Python+NumPy programs, 2018.

[12] Coralia Cartis and Katya Scheinberg. Global convergence rate analysis of unconstrained optimization methods based on probabilistic models. *Mathematical Programming*, 169(2):337–375, 2017.

[13] Coralia Cartis, Nicholas I. Gould, and Philippe L. Toint. Adaptive cubic regularisation methods for unconstrained optimization. part I: Motivation, convergence and numerical results. *Math. Program.*, 127(2):245–295, 2011.

[14] Coralia Cartis, Nicholas I M Gould, and Philippe L Toint. *Evaluation complexity of algorithms for nonconvex optimization*. MOS-SIAM Series on Optimization. Society for Industrial & Applied Mathematics, New York, NY, April 2022.

[15] R. Chen, M. Menickelly, and K. Scheinberg. Stochastic optimization using a trust-region method and random models. *Mathematical Programming*, 169(2):447–487, 2017.

[16] Alexandre Défossez, Leon Bottou, Francis Bach, and Nicolas Usunier. A simple convergence proof of adam and adagrad. *Transactions on Machine Learning Research*, 2022.

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

[18] John Duchi, Elad Hazan, and Yoram Singer. Adaptive subgradient methods for online learning and stochastic optimization. *Journal of Machine Learning Research*, 12(61):2121–2159, 2011.

[19] 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 of *Proceedings of Machine Learning Research*, pages 313–355, 2022.

[20] Matthew Faw, Litu Rout, Constantine Caramanis, and Sanjay Shakkottai. Beyond uniform smoothness: A stopped analysis of adaptive SGD. In *Proceedings of Thirty Sixth Conference on Learning Theory*, volume 195 of *Proceedings of Machine Learning Research*, pages 89–160, 2023.

[21] Sébastien Gadat and Ioana Gavra. Asymptotic study of stochastic adaptive algorithms in non-convex landscape. *Journal of Machine Learning Research*, 23(228):1–54, 2022.

[22] Saeed Ghadimid and Guanghui Lan. Stochastic First- and Zeroth-Order Methods for Nonconvex Stochastic Programming. *SIAM Journal on Optimization*, 4(23):2341–2368, 2013.

[23] Igor Gitman and Boris Ginsburg. Comparison of batch normalization and weight normalization algorithms for the large-scale image classification, 2017.

[24] Serge Gratton, Annick Sartenaer, and Philippe L. Toint. Recursive trust-region methods for multiscale nonlinear optimization. *SIAM Journal on Optimization*, 19(1):414–444, 2008.

[25] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 770–778, 2016.

[26] Tom Hennigan, Trevor Cai, Tamara Norman, Lena Martens, and Igor Babuschkin. Haiku: Sonnet for JAX, 2020.

[27] Yusu Hong and Junhong Lin Revisiting Convergence of AdaGrad with Relaxed Assumptions. In *40th Conference on Uncertainty in Artificial Intelligence (UAI)*, 1727–1750, 2024.

[28] Ali Kavis, Kfir Yehuda Levy, and Volkan Cevher. High probability bounds for a class of nonconvex algorithms with adagrad stepsize. In *The Tenth International Conference on Learning Representations, ICLR 2022, Virtual Event, April 25-29, 2022*. OpenReview.net, 2022.

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

[30] Xiaoyu Li and Francesco Orabona. On the convergence of stochastic gradient descent with adaptive stepsizes, 2018.

[31] Zijian Liu and Ta Duy Nguyen and Thien Hang Nguyen and Alina Ene and Huy L. Nguyen. High probability convergence of stochastic gradient methods In *Proceedings of the 40th International Conference on Machine Learning*, 2023.

[32] Yuxing Liu, Rui Pan, and Tong Zhang AdaGrad under Anisotropic Smoothness, arXiv:2406.15244, 2024.

[33] Ilya Loshchilov and Frank Hutter. SGDR: stochastic gradient descent with warm restarts. In *5th International Conference on Learning Representations, ICLR 2017, Toulon, France, April 24-26, 2017, Conference Track Proceedings*. OpenReview.net, 2017.

[34] Yu. Nesterov. *Introductory Lectures on Convex Optimization*. Applied Optimization. Kluwer Academic Publishers, Dordrecht, The Netherlands, 2004.

[35] Yurii Nesterov and B.T. Polyak. Cubic regularization of newton method and its global performance. *Mathematical Programming*, 108(1):177–205, 2006.

[36] Courtney Paquette and Katya Scheinberg. A stochastic line search method with expected complexity analysis. *SIAM Journal on Optimization*, 30(1):349–376, 2020.

[37] 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.

[38] Sashank J. Reddi, Satyen Kale, and Sanjiv Kumar. On the convergence of adam and beyond. In *6th International Conference on Learning Representations, ICLR 2018, Vancouver, BC, Canada, April 30 - May 3, 2018, Conference Track Proceedings*, 2018.

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

[40] Bohan Wang, Huishuai Zhang, Zhiming Ma, and Wei Chen. Convergence of adagrad for non-convex objectives: Simple proofs and relaxed assumptions. In *Proceedings of Thirty Sixth Conference on Learning Theory*, volume 195, pages 161–190, 2023.

[41] R. Ward, X. Wu, and L. Bottou. AdaGrad stepsizes: Sharp convergence over nonconvex landscapes. In K. Chaudhuri and R. Salakhutdinov, editors, *Proceedings of the 36th International Conference on Machine Learning*, volume 97, pages 6677–6686, 2019.

[42] Sergey Zagoruyko and Nikos Komodakis. Wide residual networks. In *Proceedings of the British Machine Vision Conference 2016, BMVC 2016, York, UK, September 19-22, 2016*. BMVA Press, 2016.

[43] Manzil Zaheer, Sashank Reddi, Devendra Sachan, Satyen Kale, and Sanjiv Kumar. Adaptive methods for nonconvex optimization. In *Advances in Neural Information Processing Systems*, volume 31, 2018.

[44] Jingzhao Zhang, Tianxing He, Suvrit Sra, and Ali Jadbabaie. Why gradient clipping accelerates training: A theoretical justification for adaptivity. In *International Conference on Learning Representations*, 2020.

[45] Dongruo Zhou, Jinghui Chen, Yuan Cao, Ziyan Yang, and Quanquan Gu. On the convergence of adaptive gradient methods for nonconvex optimization. *Transactions on Machine Learning Research*, 2024. Featured Certification.

[46] Fangyu Zou, Li Shen, Zequn Jie, Weizhong Zhang, and Wei Liu. A sufficient condition for convergences of adam and rmsprop. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2019, Long Beach, CA, USA, June 16-20, 2019*, pages 11127–11135. Computer Vision Foundation / IEEE, 2019.

A first technical lemma

Lemma A.1. *Let $\mu \in (0,1]$. Let x, $y \in \mathbb{R}^+ \setminus \{0\}$. Then*

$$\frac{|x^\mu - y^\mu|}{x^\mu y^\mu} \leq \mu \frac{|x-y|}{xy^\mu} + \mu \frac{|x-y|}{x^\mu y}. \tag{A.1}$$

Proof. Let us first consider the case $x \geq y$. Remembering that $u^\mu \leq 1 + \mu(u-1)$ for $u > 0$ and taking $u = \frac{x}{y}$, we successively derive that

$$\begin{aligned} \frac{x^\mu}{y^\mu} &\leq 1 + \mu\left(\frac{x}{y} - 1\right), \\ x^\mu - y^\mu &\leq \mu\left(\frac{xy^\mu}{y} - y^\mu\right) = \mu y^{\mu-1}(x-y), \\ \frac{x^\mu - y^\mu}{x^\mu y^\mu} &\leq \mu \frac{x-y}{x^\mu y}. \end{aligned} \tag{A.2}$$

Hence the inequality (A.1) is valid when $x \geq y$. For the symmetric case ($y \geq x$), we similarly obtain that

$$\frac{y^\mu - x^\mu}{x^\mu y^\mu} \leq \mu \frac{y-x}{y^\mu x}. \tag{A.3}$$

Combining (A.2) and (A.3) yields the desired result.

Proof of Lemma 3.3

Let us consider an iteration index $j \geq 0$ and a component index $i \in \{1, \ldots, n\}$. We first use the definition of s^L in (2.4) and the fact that $w_{i,j} = v_{i,j}^\mu$ (Assumption 3.1) to obtain that

$$\mathbb{E}_j\left[\gamma_j G_{i,j} s_{i,j}^L\right] = -\mathbb{E}_j\left[\gamma_j \frac{G_{i,j} g_{i,j}}{v_{i,j}^\mu}\right] = -\mathbb{E}_j\left[\gamma_j \frac{G_{i,j} g_{i,j}}{\mathbb{E}_j[v_{i,j}]^\mu}\right] + \mathbb{E}_j\left[\gamma_j G_{i,j} g_{i,j}\left(\frac{1}{\mathbb{E}_j[v_{i,j}]^\mu} - \frac{1}{v_{i,j}^\mu}\right)\right]. \tag{A.1}$$

Using that $G_{i,j}$ and γ_j are mesurable with respect to the past and Assumption 2.4, we derive that,

$$\mathbb{E}_j\left[-\frac{\gamma_j G_{i,j} g_{i,j}}{\mathbb{E}_j[v_{i,j}]^\mu}\right] = -\frac{\gamma_j G_{i,j}}{\mathbb{E}_j[v_{i,j}]^\mu}\mathbb{E}_j[g_{i,j}] = -\frac{\gamma_j G_{i,j}^2}{\mathbb{E}_j[v_{i,j}]^\mu} \leq -\gamma_{\text{low}} \frac{G_{i,j}^2}{\mathbb{E}_j[v_{i,j}]^\mu}, \tag{A.2}$$

where we used the measurability of $\mathbb{E}_j[v_{i,j}]^\mu$ with respect to the past. Combining (A.1) and (A.2) gives that

$$\mathbb{E}_j\left[\gamma_j G_{i,j} s_{i,j}^L\right] \leq -\gamma_{\text{low}} \frac{G_{i,j}^2}{\mathbb{E}_j[v_{i,j}]^\mu} + \mathbb{E}_j\left[\underbrace{\gamma_j G_{i,j} g_{i,j} \frac{v_{i,j}^\mu - \mathbb{E}_j[v_{i,j}]^\mu}{v_{i,j}^\mu \mathbb{E}_j[v_{i,j}]^\mu}}_{A}\right]. \tag{A.3}$$

We now derive an upper bound on the absolute value of the A term by successively using Lemma (A.1), Assumption 3.1 and the bound $\gamma_j \leq 1$ to obtain that

$$|A| = |\gamma_j G_{i,j} g_{i,j}| \frac{|v_{i,j}^{\mu} - \mathbb{E}_j[v_{i,j}]^{\mu}|}{v_{i,j}^{\mu}\mathbb{E}_j[v_{i,j}]^{\mu}} \leq \mu|\gamma_j G_{i,j} g_{i,j}| \frac{|v_{i,j} - \mathbb{E}_j[v_{i,j}]|}{v_{i,j}^{\mu}\mathbb{E}_j[v_{i,j}]} + \mu|\gamma_j G_{i,j} g_{i,j}| \frac{|v_{i,j} - \mathbb{E}_j[v_{i,j}]|}{v_{i,j}\mathbb{E}_j[v_{i,j}]^{\mu}}$$

$$\leq \mu \underbrace{|G_{i,j} g_{i,j}| \kappa_v \frac{\mathbb{E}_j\left[g_{i,j}^2\right]}{v_{i,j}^{\mu}\mathbb{E}_j[v_{i,j}]}}_{B} + \mu \underbrace{|G_{i,j} g_{i,j}| \kappa_v \frac{g_{i,j}^2}{v_{i,j}^{\mu}\mathbb{E}_j[v_{i,j}]}}_{C}$$

$$+ \mu \underbrace{|G_{i,j} g_{i,j}| \kappa_v \frac{\mathbb{E}_j\left[g_{i,j}^2\right]}{v_{i,j}\mathbb{E}_j[v_{i,j}]^{\mu}}}_{D} + \mu \underbrace{|G_{i,j} g_{i,j}| \kappa_v \frac{g_{i,j}^2}{v_{i,j}\mathbb{E}_j[v_{i,j}]^{\mu}}}_{E}.$$

We now use Young's inequality with $p = q = 2$, that is

$$\forall \lambda > 0, x,\, y \in \mathbb{R}^+,\, xy \leq \frac{\lambda}{2} x^2 + \frac{y^2}{2\lambda}, \tag{A.4}$$

to successively handle the four terms in the last bound.

- For the first term B, we choose

$$x = \frac{|G_{i,j}|}{\mathbb{E}_j[v_{i,j}]^{\mu}}, \quad \lambda = \frac{\gamma_{\text{low}}\mathbb{E}_j[v_{i,j}]^{\mu}}{4} \text{ and } y = \kappa_v |g_{i,j}| \frac{\mathbb{E}_j\left[g_{i,j}^2\right]}{v_{i,j}^{\mu}\mathbb{E}_j[v_{i,j}]^{1-\mu}}.$$

Using (A.4), Assumptions 2.4, 3.2 and (3.3), we obtain that

$$\begin{aligned} B &\leq \gamma_{\text{low}} \frac{G_{i,j}^2}{8\mathbb{E}_j[v_{i,j}]^{\mu}} + 2\frac{\kappa_v^2}{\gamma_{\text{low}}} \frac{g_{i,j}^2}{v_{i,j}^{2\mu}} \frac{\mathbb{E}_j\left[g_{i,j}^2\right]^2}{\mathbb{E}_j[v_{i,j}]^{2-\mu}}, \\ &\leq \gamma_{\text{low}} \frac{G_{i,j}^2}{8\mathbb{E}_j[v_{i,j}]^{\mu}} + 2\frac{\kappa_v^2}{\gamma_{\text{low}}} \mathbb{E}_j\left[g_{i,j}^2\right]^{\mu} \frac{g_{i,j}^2}{v_{i,j}^{2\mu}} \\ &\leq \gamma_{\text{low}} \frac{G_{i,j}^2}{8\mathbb{E}_j[v_{i,j}]^{\mu}} + 2\frac{\kappa_v^2}{\gamma_{\text{low}}} \kappa_g^{2\mu} \frac{g_{i,j}^2}{v_{i,j}^{2\mu}}. \end{aligned}$$

Taking now the expectation over $\mathbb{E}_j[.]$ yields that

$$\mathbb{E}_j[B] \leq \gamma_{\text{low}} \frac{G_{i,j}^2}{8\mathbb{E}_j[v_{i,j}]^{\mu}} + 2\frac{\kappa_v^2}{\gamma_{\text{low}}} \kappa_g^{2\mu} \mathbb{E}_j\left[\frac{g_{i,j}^2}{w_{i,j}^2}\right]. \tag{A.5}$$

- Now consider the C term. In this case, we choose

$$x = \frac{|G_{i,j} g_{i,j}|}{\mathbb{E}_j[v_{i,j}]^{\mu}}, \quad \lambda = \gamma_{\text{low}} \frac{\mathbb{E}_j[v_{i,j}]^{\mu}}{4\mathbb{E}_j\left[g_{i,j}^2\right]} \text{ and } y = \kappa_v \frac{g_{i,j}^2}{v_{i,j}^{\mu}\mathbb{E}_j[v_{i,j}]^{1-\mu}}$$

to deduce from (A.4) that

$$\begin{aligned} C &\leq \gamma_{\text{low}} \frac{G_{i,j}^2}{8\mathbb{E}_j[v_{i,j}]^\mu} \frac{g_{i,j}^2}{\mathbb{E}_j\left[g_{i,j}^2\right]} + 2\frac{\kappa_v^2}{\gamma_{\text{low}}} \frac{g_{i,j}^4}{v_{i,j}^{2\mu}} \frac{\mathbb{E}_j\left[g_{i,j}^2\right]}{\mathbb{E}_j[v_{i,j}]^{2-\mu}} \\ &\leq \gamma_{\text{low}} \frac{G_{i,j}^2}{8\mathbb{E}_j[v_{i,j}]^\mu} \frac{g_{i,j}^2}{\mathbb{E}_j\left[g_{i,j}^2\right]} + 2\frac{\kappa_v^2}{\gamma_{\text{low}}} \kappa_g^2 \frac{g_{i,j}^2}{v_{i,j}^{2\mu}} \frac{1}{\mathbb{E}_j[v_{i,j}]^{1-\mu}} \\ &\leq \gamma_{\text{low}} \frac{G_{i,j}^2}{8\mathbb{E}_j[v_{i,j}]^\mu} \frac{g_{i,j}^2}{\mathbb{E}_j\left[g_{i,j}^2\right]} + 2\frac{\kappa_v^2}{\gamma_{\text{low}}} \frac{\kappa_g^2}{\varsigma_{\min}^{1-\mu}} \frac{g_{i,j}^2}{v_{i,j}^{2\mu}}, \end{aligned}$$

where we successively used the facts that $\mathbb{E}_j\left[g_{i,j}^2\right] \leq \mathbb{E}_j[v_{i,j}]$ (because of (3.3)), $g_{i,j}^2 \leq \kappa_g^2$ (because of Assumption 2.4) and $\mathbb{E}_j[v_{i,j}]^{1-\mu} \geq \varsigma_{\min}^{1-\mu}$ (because of (3.2)). Taking the expectation over $\mathbb{E}_j[.]$ then gives that

$$\mathbb{E}_j[C] \leq \gamma_{\text{low}} \frac{G_{i,j}^2}{8\mathbb{E}_j[v_{i,j}]^\mu} + 2\frac{\kappa_v^2}{\gamma_{\text{low}}} \frac{\kappa_g^2}{\varsigma_{\min}^{1-\mu}} \mathbb{E}_j\left[\frac{g_{i,j}^2}{w_{i,j}^2}\right]. \tag{A.6}$$

(Note that we can divide by $\mathbb{E}_j\left[g_{i,j}^2\right]$ above, as it suffice to notice that $\mathbb{E}_j\left[g_{i,j}^2\right] = 0$ implies $g_{i,j}^2 = 0$. C would then be equal to zero and (A.7) would still be verified.)

• Let us now handle the D term. Choosing

$$x = \frac{|G_{i,j}|}{\mathbb{E}_j[v_{i,j}]^\mu}, \quad \lambda = \gamma_{\text{low}} \frac{\mathbb{E}_j[v_{i,j}]^\mu}{4} \text{ and } y = \kappa_v |g_{i,j}| \frac{\mathbb{E}_j\left[g_{i,j}^2\right]}{v_{i,j}},$$

we now deduce from (A.4) that

$$\begin{aligned} D &\leq \gamma_{\text{low}} \frac{G_{i,j}^2}{8\mathbb{E}_j[v_{i,j}]^\mu} + 2\frac{\kappa_v^2}{\gamma_{\text{low}}} \frac{g_{i,j}^2 \mathbb{E}_j\left[g_{i,j}^2\right]^2}{\mathbb{E}_j[v_{i,j}]^\mu v_{i,j}^2}, \\ &\leq \gamma_{\text{low}} \frac{G_{i,j}^2}{8\mathbb{E}_j[v_{i,j}]^\mu} + 2\frac{\kappa_v^2}{\gamma_{\text{low}}} \frac{g_{i,j}^2}{v_{i,j}^{2\mu}} \frac{1}{v_{i,j}^{2-2\mu}} \frac{\mathbb{E}_j\left[g_{i,j}^2\right]^2}{\mathbb{E}_j[v_{i,j}]^\mu} \\ &\leq \gamma_{\text{low}} \frac{G_{i,j}^2}{8\mathbb{E}_j[v_{i,j}]^\mu} + 2\frac{\kappa_v^2}{\gamma_{\text{low}}} \frac{g_{i,j}^2}{v_{i,j}^{2\mu}} \frac{1}{v_{i,j}^{2-2\mu}} \mathbb{E}_j\left[g_{i,j}^2\right]^{2-\mu} \\ &\leq \gamma_{\text{low}} \frac{G_{i,j}^2}{8\mathbb{E}_j[v_{i,j}]^\mu} + 2\frac{\kappa_v^2}{\gamma_{\text{low}}} \frac{\kappa_g^{4-2\mu}}{\varsigma_{\min}^{2-2\mu}} \frac{g_{i,j}^2}{v_{i,j}^{2\mu}}, \end{aligned}$$

where, as for the C term, we used the facts that $\mathbb{E}_j\left[g_{i,j}^2\right]^\mu \leq \mathbb{E}_j[v_{i,j}]^\mu$, $g_{i,j}^2 \leq \kappa_g^2$ and $v_{i,j}^{2-2\mu} \geq \varsigma_{\min}^{2-2\mu}$. Taking the expectation $\mathbb{E}_j[.]$ yields, in this case, that

$$\mathbb{E}_j[D] \leq \gamma_{\text{low}} \frac{G_{i,j}^2}{8\mathbb{E}_j[v_{i,j}]^\mu} + 2\frac{\kappa_v^2}{\gamma_{\text{low}}} \frac{\kappa_g^{4-2\mu}}{\varsigma_{\min}^{2-2\mu}} \mathbb{E}_j\left[\frac{g_{i,j}^2}{w_{i,j}^2}\right]. \tag{A.7}$$

• Finally consider the E term. Choosing

$$x = \frac{|G_{i,j} g_{i,j}|}{\mathbb{E}_j[v_{i,j}]^\mu}, \quad \lambda = \gamma_{\text{low}} \frac{\mathbb{E}_j[v_{i,j}]^\mu}{4\mathbb{E}_j\left[g_{i,j}^2\right]} \text{ and } y = \kappa_v \frac{g_{i,j}^2}{v_{i,j}}$$

in (A.4) then gives that

$$\begin{aligned}
\mathbb{E}_j[E] &\leq \gamma_{\text{low}} \frac{G_{i,j}^2}{8\mathbb{E}_j[v_{i,j}]^\mu} \frac{g_{i,j}^2}{\mathbb{E}_j\left[g_{i,j}^2\right]} + 2\frac{\kappa_v^2}{\gamma_{\text{low}}} \frac{g_{i,j}^4 \mathbb{E}_j\left[g_{i,j}^2\right]}{\mathbb{E}_j[v_{i,j}]^\mu \, v_{i,j}^2} \\
&\leq \gamma_{\text{low}} \frac{G_{i,j}^2}{8\mathbb{E}_j[v_{i,j}]^\mu} \frac{g_{i,j}^2}{\mathbb{E}_j\left[g_{i,j}^2\right]} + 2\frac{\kappa_v^2}{\gamma_{\text{low}}} \mathbb{E}_j\left[g_{i,j}^2\right]^{1-\mu} \frac{g_{i,j}^2}{v_{i,j}^{2\mu}} \left(\frac{1}{v_{i,j}^{1-2\mu}} \frac{g_{i,j}^2}{v_{i,j}} \mathbb{1}_{\mu<\frac{1}{2}} + \frac{|g_{i,j}^{4-4\mu}|}{v_{i,j}^{2-2\mu}} |g_{i,j}^{4\mu-2}| \mathbb{1}_{\mu\geq\frac{1}{2}} \right) \\
&\leq \gamma_{\text{low}} \frac{G_{i,j}^2}{8\mathbb{E}_j[v_{i,j}]^\mu} \frac{g_{i,j}^2}{\mathbb{E}_j\left[g_{i,j}^2\right]} + 2\frac{\kappa_v^2}{\gamma_{\text{low}}} \kappa_g^{2-2\mu} \frac{g_{i,j}^2}{v_{i,j}^{2\mu}} \left(\frac{1}{\varsigma_{\min}^{1-2\mu}} \mathbb{1}_{\mu<\frac{1}{2}} + \kappa_g^{4\mu-2} \mathbb{1}_{\mu\geq\frac{1}{2}} \right),
\end{aligned}$$

where we once more used the facts that $\mathbb{E}_j\left[g_{i,j}^2\right]^\mu \leq \mathbb{E}_j[v_{i,j}]^\mu$ and $|g_{i,j}| \leq \kappa_g$, in turn implying that

$$g_{i,j}^2 \leq v_{i,j} \text{ and } v_{i,j} \geq \varsigma_{\min} \quad \text{if} \quad \mu < \tfrac{1}{2}$$

and

$$|g_{i,j}^{4-4\mu}| \leq v_{i,j}^{2-2\mu} \text{ and } |g_{i,j}^{4\mu-2}| \leq \kappa_g^{4\mu-2} \quad \text{if} \quad \mu \geq \tfrac{1}{2}.$$

Taking the expectation $\mathbb{E}_j[.]$, we deduce that

$$\mathbb{E}_j[E] \leq \gamma_{\text{low}} \frac{G_{i,j}^2}{8\mathbb{E}_j[v_{i,j}]^\mu} + \frac{\kappa_v^2}{\gamma_{\text{low}}} \kappa_g^{2-2\mu} \left(\frac{1}{\varsigma_{\min}^{1-2\mu}} \mathbb{1}_{\mu<\frac{1}{2}} + \kappa_g^{4\mu-2} \mathbb{1}_{\mu\geq\frac{1}{2}} \right) \mathbb{E}_j\left[\frac{g_{i,j}^2}{w_{i,j}^2} \right]. \tag{A.8}$$

Summing now (A.6), (A.7), (A.7) and (A.8) and substituting the obtained upper-bound of A in (A.3), we finally obtain (3.4) with (3.5).

Proof of Lemma 3.4

Consider first the case where $\alpha \neq 1$ and note that $\frac{1}{(1-\alpha)} x^{1-\alpha}$ is then a non-decreasing and concave function on $(0, +\infty)$. Setting $b_{-1} = 0$ and using these properties, we obtain that, for $j \geq 0$,

$$\begin{aligned}
\frac{a_j}{(\varsigma + b_j)^\alpha} &\leq \frac{1}{1-\alpha} \left((\varsigma + b_j)^{1-\alpha} - (\varsigma + b_j - a_j)^{1-\alpha} \right) \\
&\leq \frac{1}{1-\alpha} \left((\varsigma + b_j)^{1-\alpha} - (\varsigma + b_{j-1})^{1-\alpha} \right).
\end{aligned}$$

We then obtain (3.6) by summing this inequality for $j \in \{0, \ldots, k\}$.

Suppose now that $\alpha = 1$, We then use the concavity and non-decreasing nature of the logarithm to derive that

$$\frac{a_j}{(\varsigma + b_j)^\alpha} = \frac{a_j}{(\varsigma + b_j)} \leq \log(\varsigma + b_j) - \log(\varsigma + b_j - a_j) \leq \log(\varsigma + b_j) - \log(\varsigma + b_{j-1}).$$

The inequality (3.7) then again follows by summing for $j \in \{0, \ldots, k\}$.

Proof of Lemma 4.1

We have that

$$\begin{aligned}\mathbb{E}_j\left[-\gamma_j\frac{G_{i,j}g_{i,j}}{w_{i,j}}\right] &\leq -\gamma_{\text{low}}\frac{G_{i,j}^2}{\mathbb{E}_j[w_{i,j}]} + \mathbb{E}_j\left[|\gamma_j G_{i,j}g_{i,j}|\frac{|w_{i,j}-\mathbb{E}_j[w_{i,j}]\,|}{w_{i,j}\mathbb{E}_j[w_{i,j}]}\right]\\ &= -\gamma_{\text{low}}\frac{G_{i,j}^2}{\mathbb{E}_j[w_{i,j}]} + \mathbb{E}_j\left[|\gamma_j G_{i,j}g_{i,j}|\frac{|\xi_{i,j}-\mathbb{E}_j[\xi_{i,j}]\,|}{\xi_{i,j}(j+1)^\mu\mathbb{E}_j[\xi_{i,j}]}\right]\\ &\leq -\gamma_{\text{low}}\frac{G_{i,j}^2}{\mathbb{E}_j[w_{i,j}]} + \mathbb{E}_j[A] \qquad \text{(A.1)}\end{aligned}$$

where

$$A = |\gamma_j G_{i,j}g_{i,j}|\,\frac{|\xi_{i,j}-\mathbb{E}_j[\xi_{i,j}]\,|}{\xi_{i,j}(j+1)^\mu\mathbb{E}_j[\xi_{i,j}]} \qquad \text{(A.2)}$$

In order to compute a bound on the right-hand side of (A.1), we consider the *maxgi* and *avrgi* cases separately. Consider the *maxgi* case first. We see that

$$\begin{aligned}\mathbb{E}_j[|\xi_{i,j}-\mathbb{E}_j[\xi_{i,j}]\,|] &= \mathbb{E}_j\left[\left|\mathbb{1}_{\mathcal{A}_j}\xi_{i,j}-\mathbb{E}_j\left[\mathbb{1}_{\mathcal{A}_j}\xi_{i,j}\right]+\mathbb{1}_{\mathcal{A}_j^C}\xi_{i,j}-\mathbb{E}_j\left[\mathbb{1}_{\mathcal{A}_j^C}\xi_{i,j}\right]\right|\right]\\ &= \mathbb{E}_j\left[\left|\mathbb{1}_{\mathcal{A}_j}|g_{i,j}|-\mathbb{E}_j\left[\mathbb{1}_{\mathcal{A}_j}|g_{i,j}|\right]+\mathbb{1}_{\mathcal{A}_j^C}\xi_{i,j-1}-\mathbb{E}_j\left[\mathbb{1}_{\mathcal{A}_j^C}\xi_{i,j-1}\right]\right|\right]\\ &\leq \mathbb{E}_j\left[\left|\mathbb{1}_{\mathcal{A}_j}|g_{i,j}|-\mathbb{E}_j\left[\mathbb{1}_{\mathcal{A}_j}|g_{i,j}|\right]+\xi_{i,j-1}\left(\mathbb{1}_{\mathcal{A}_j^C}-\mathbb{E}_j\left[\mathbb{1}_{\mathcal{A}_j^C}\right]\right)\right|\right]\\ &\leq \mathbb{E}_j\left[\mathbb{1}_{\mathcal{A}_j}|g_{i,j}|+\mathbb{E}_j\left[\mathbb{1}_{\mathcal{A}_j}|g_{i,j}|\right]+\xi_{i,j-1}\left|\mathbb{1}_{\mathcal{A}_j^C}-\mathbb{E}_j\left[\mathbb{1}_{\mathcal{A}_j^C}\right]\right|\right]\\ &\leq \kappa_g(\mathbb{E}_j\left[\mathbb{1}_{\mathcal{A}_j}+\mathbb{E}_j\left[\mathbb{1}_{\mathcal{A}_j}\right]\right])+\kappa_g\mathbb{E}_j\left[\left|\mathbb{1}_{\mathcal{A}_j^C}-\mathbb{E}_j\left[\mathbb{1}_{\mathcal{A}_j^C}\right]\right|\right]\\ &\leq 2\kappa_g\mathbb{P}_j[\mathcal{A}_j]+\kappa_g\mathbb{E}_j\left[\left|\mathbb{1}_{\mathcal{A}_j^C}-\mathbb{E}_j\left[\mathbb{1}_{\mathcal{A}_j^C}\right]\right|\right], \qquad \text{(A.3)}\end{aligned}$$

where we used the bound $\xi_{i,j} \leq \kappa_g$ for all j (resulting from Assumption 2.4) to obtain the penultimate inequality. Now

$$\mathbb{E}_j\left[\left|\mathbb{1}_{A_j^C}-\mathbb{E}_j\left[\mathbb{1}_{A_j^C}\right]\right|\right] = \mathbb{E}_j\left[|1-\mathbb{1}_{A_j}-\mathbb{E}_j\left[1-\mathbb{1}_{A_j}\right]|\right] = \mathbb{E}_j\left[|\mathbb{1}_{A_j}-\mathbb{E}_j\left[\mathbb{1}_{A_j}\right]|\right] \leq 2\,\mathbb{P}_j(A_j).$$

We then obtain (4.3)-(4.4) by substituting this last inequality in (A.3) and combing the result with the bound

$$\frac{|\gamma_j G_{i,j}g_{i,j}|}{\xi_{i,j}(j+1)^\mu\mathbb{E}_j[\xi_{i,j}]} \leq \frac{\kappa_g^2}{\varsigma^2(j+1)^\mu},$$

(A.1) and (A.2).

Now consider the *avrgi* case. Analogously to (A.3) and using the identity $\max(a,b) = \frac{1}{2}(a+$

$b + |a-b|$), we deduce from (4.2) that

$$
\begin{aligned}
|\mathbb{E}_j[\xi_{i,j}] - \xi_{i,j}| &= \frac{1}{2}\left|\mathbb{E}_j\left[\varsigma + \frac{1}{j+1}\sum_{k=0}^{j}|g_{i,k}| + |\frac{1}{j+1}\sum_{k=0}^{j}|g_{i,k}| - \varsigma|\right]\right. \\
&\qquad \left. -\varsigma - \frac{1}{j+1}\sum_{k=0}^{j}|g_{i,k}| - |\frac{1}{j+1}\sum_{k=0}^{j}|g_{i,k}| - \varsigma|\right| \\
&= \frac{1}{2}\left|\frac{1}{(j+1)}\mathbb{E}_j[|g_{i,j}|] + \mathbb{E}_j\left[|\frac{1}{j+1}\sum_{k=0}^{j}|g_{i,k}| - \varsigma|\right] - \frac{1}{(j+1)}|g_{i,j}| - |\frac{1}{j+1}\sum_{k=0}^{j}|g_{i,k}| - \varsigma|\right| \\
&\leq \frac{1}{2(j+1)}\left|\mathbb{E}_j[|g_{i,j}|] - |g_{i,j}|\right| \\
&\quad + \frac{1}{2}\left|\frac{1}{(j+1)}\mathbb{E}_j[|g_{i,j}|] + |\frac{1}{j+1}\sum_{k=0}^{j-1}|g_{i,k}| - \varsigma| + \frac{1}{(j+1)}|g_{i,j}| - |\frac{1}{j+1}\sum_{k=0}^{j-1}|g_{i,k}| - \varsigma|\right| \\
&\leq \frac{1}{(j+1)}\left|\mathbb{E}_j[|g_{i,j}|] + |g_{i,j}|\right|,
\end{aligned}
$$

where we used that $|\frac{1}{j+1}\sum_{k=0}^{j-1}|g_{i,k}| - \varsigma|$ is mesurable with respect to the past. This inequality, the definition (A.2) and the bounds $\gamma_j \leq 1$ and $(j+1)^{\mu/2} \leq j+1$ then give that

$$
A \leq \underbrace{\frac{|G_{i,j}|g_{i,j}^2}{(j+1)^{\mu/2+\mu}\mathbb{E}_j[\xi_{i,j}]\,\xi_{i,j}}}_{B} + \underbrace{\frac{|G_{i,j}g_{i,j}\mathbb{E}_j[|g_{i,j}|]\,|}{(j+1)^{\mu/2+\mu}\mathbb{E}_j[\xi_{i,j}]\,\xi_{i,j}}}_{C}. \tag{A.4}
$$

We now use Young's inequality with $p = q = 2$, that is

$$
\forall \lambda > 0, x,\, y \in \mathbb{R}^+,\; xy \leq \frac{\lambda}{2}x^2 + \frac{y^2}{2\lambda}, \tag{A.5}
$$

to successively handle the two terms of (A.4).

- For the first term B, we choose

$$
x = \frac{|G_{i,j}|}{(j+1)^{\mu/2}\sqrt{\mathbb{E}_j[\xi_{i,j}]}}, \quad \lambda = \frac{\gamma_{\text{low}}}{2} \;\text{ and }\; y = \frac{g_{i,j}^2}{\sqrt{\mathbb{E}_j[\xi_{i,j}]}\xi_{i,j}(j+1)^{\mu}}
$$

yielding

$$
B \leq \frac{\gamma_{\text{low}}G_{i,j}^2}{4(j+1)^{\mu}\mathbb{E}_j[\xi_{i,j}]} + \frac{1}{\gamma_{\text{low}}}\frac{g_{i,j}^4}{\mathbb{E}_j[\xi_{i,j}]\,\xi_{i,j}^2(j+1)^{2\mu}} \leq \frac{\gamma_{\text{low}}G_{i,j}^2}{4(j+1)^{\mu}\mathbb{E}_j[\xi_{i,j}]} + \frac{1}{\gamma_{\text{low}}}\frac{g_{i,j}^2\kappa_g^2}{\varsigma\xi_{i,j}^2(j+1)^{2\mu}},
$$

where we used that $|g_{i,j}| \leq \kappa_g$ and $\xi_{i,j} \geq \varsigma$. Taking now $\mathbb{E}_j[\cdot]$ in the previous inequality, using that $w_{i,j} = \xi_{i,j}(j+1)^{\mu}$, we derive that

$$
\mathbb{E}_j[B] \leq \frac{\gamma_{\text{low}}G_{i,j}^2}{4\mathbb{E}_j[w_{i,j}]} + \frac{\kappa_g^2}{\varsigma\gamma_{\text{low}}}\mathbb{E}_j\left[\frac{g_{i,j}^2}{w_{i,j}^2}\right] \tag{A.6}
$$

- Now consider the C term. Again, Young's inequality with

$$
x = \frac{|G_{i,j}|}{(j+1)^{\mu/2}\sqrt{\mathbb{E}_j[\xi_{i,j}]}}, \quad \lambda = \frac{\gamma_{\text{low}}}{2} \;\text{ and }\; y = \frac{|g_{i,j}|\mathbb{E}_j[|g_{i,j}|]}{\sqrt{\mathbb{E}_j[\xi_{i,j}]}\xi_{i,j}(j+1)^{\mu}}
$$

yields that

$$\mathbb{E}_j[C] \leq \frac{\gamma_{\text{low}} G_{i,j}^2}{4\mathbb{E}_j[w_{i,j}]} + \frac{\kappa_g^2}{\varsigma\gamma_{\text{low}}}\mathbb{E}_j\left[\frac{g_{i,j}^2}{w_{i,j}^2}\right], \tag{A.7}$$

where we used that $|g_{i,j}| \leq \kappa_g$ and $\xi_{i,j} \geq \varsigma$. Taking $\mathbb{E}_j[.]$ in (A.4), using (A.7) and (A.6) and injecting the obtained bound into (A.1), we obtain (4.3) with (4.5).