The stochastic Ravine accelerated gradient method with general extrapolation coefficients

Hedy Attouch, Jalal M. Fadili, Vyacheslav Kungurtsev

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

The Stochastic Ravine Accelerated Gradient Method with General Extrapolation Coefficients

Hedy Attouch, Jalal Fadili and Vyacheslav Kungurtsev

Received: date / Accepted: date*

Abstract In a real Hilbert space domain setting, we study the convergence properties of the stochastic Ravine accelerated gradient method for convex differentiable optimization. We consider the general form of this algorithm where the extrapolation coefficients can vary with each iteration, and where the evaluation of the gradient is subject to random errors. This general treatment models a breadth of practical algorithms and numerical implementations. We show that, under a proper tuning of the extrapolation parameters, and when the error variance associated with the gradient evaluations or the step-size sequences vanish sufficiently fast, the Ravine method provides fast convergence of the values both in expectation and almost surely. We also improve the convergence rates from $\mathcal{O}(\cdot)$ to $o(\cdot)$ in expectation and almost sure sense. Moreover, we show almost sure summability property of the gradients, which implies the fast convergence of the gradients towards zero. This property reflects the fact that the high-resolution ODE of the Ravine method includes a Hessian-driven damping term. When the space is also separable, our analysis allows to establish almost sure weak convergence of the sequence of iterates provided by the algorithm. We finally specialize the analysis to consider different parameter choices, including vanishing and constant (heavy ball method with friction) damping parameter, and present a

* The insight and motivation for the study of the Ravine method, in light of its resemblance to Nesterov's, as well as many of the derivations in this paper, was the work of our beloved friend and colleague Hedy Attouch. As one of the final contributions of Hedy's long and illustrious career before his unfortunate recent departure, the other authors are fortunate to have worked with him on this topic, and hope that the polished manuscript is a valuable step in honoring his legacy.

Hedy Attouch
IMAG CNRS UMR 5149,
Université Montpellier,
Place Eugène Bataillon, 34095 Montpellier CEDEX 5, France.
hedy.attouch@umontpellier.fr

Jalal Fadili
GREYC CNRS UMR 6072
Ecole Nationale Supérieure d'Ingénieurs de Caen
14050 Caen Cedex France.
Jalal.Fadili@greyc.ensicaen.fr

Vyacheslav Kungurtsev
Department of Computer Science
Faculty of Electrical Engineering
Czech Technical University,
12000 Prague, Czechia
vyacheslav.kungurtsev@fel.cvut.cz

comprehensive landscape of the tradeoffs in speed and accuracy associated with these parameter choices and statistical properties on the sequence of errors in the gradient computations. We provide a thorough discussion of the similarities and differences with the Nesterov accelerated gradient which satisfies similar asymptotic convergence rates.

Keywords Ravine method · Nesterov accelerated gradient method · general extrapolation coefficient · stochastic errors · Hessian driven damping · convergence rates · Lyapunov analysis

Mathematics Subject Classification (2020) 37N40 · 46N10 · 49M30 · 65B99 · 65K05 · 65K10 · 90B50 · 90C25

1 Introduction

Given a real Hilbert space $\mathcal{H}$, our work concerns is concerned with fast numerical resolution of the convex minimization problem

$$\min\{f(x) :\ x \in \mathcal{H}\}, \tag{$\mathcal{P}$}$$

where we make the following standing assumptions:

$$\begin{cases} f : \mathcal{H} \to \mathbb{R} \ \text{ is differentiable, } \nabla f \text{ is } L-\text{Lipschitz continuous, } S = \operatorname{argmin} f \neq \emptyset. \\ (s_k)_{k\in\mathbb{N}} \ \text{ is a positive sequence with } s_k L \in]0,1]. \end{cases} \tag{H}$$

To solve ($\mathcal{P}$), we consider the Ravine Accelerated Gradient algorithm ((RAG)$_{\gamma_k}$ for short), which generates iterates $(y_k, w_k)_{k\in\mathbb{N}}$ satisfying

$$\begin{cases} w_k = y_k - s_k \nabla f(y_k) \\ y_{k+1} = w_k + \gamma_k (w_k - w_{k-1}). \end{cases} \tag{(RAG)$_{\gamma_k}$}$$

Let us indicate the role of the different parameters involved in the above algorithm:

a) The positive parameter sequence $(s_k)_{k\in\mathbb{N}}$ is the step-size sequence applied to the gradient based update.
b) The non-negative extrapolation coefficients $(\gamma_k)_{k\in\mathbb{N}}$ are linked to the inertial character of the algorithm. They can be viewed as control parameters for optimization purposes.
c) In order to inform about the practical performance of algorithms realizing this method in common applications, we will analyze the convergence properties when the gradient terms are calculated with stochastic errors. Formally, we consider $\nabla f(y_k) + e_k$ instead of $\nabla f(y_k)$ in (RAG)$_{\gamma_k}$ where e_k is a *zero-mean stochastic noise* term.

One of the motivations for this additive perturbation model comes comes from stochastic optimization problems of the form

$$f(x) = \int_{\Xi} F(x,\xi) d\mu(\xi) := \mathbb{E}\left[F(x,\xi)\right],\ F : \mathcal{H} \times \Xi \to \mathbb{R} \tag{1}$$

where $(\Xi, \mathcal{F}, \mu)$ is a probability space, $F(x,\cdot)$ is μ-integrable for any $x \in \mathcal{H}$, and $F(\cdot,\xi) \in C^1(\mathcal{H})$ for any ξ. Problem (1) is very popular in many applications including machine learning and signal processing. As computing $\nabla f(x)$ is computationally very expensive or even impossible,

the popular alternative is to draw m independent samples of ξ, say $(\xi_i)_{1\leq i\leq m}$, and compute the empirical average estimate

$$\widehat{\nabla f}(x) = \frac{1}{m}\sum_{i=1}^{m} \nabla F(x,\xi_i). \tag{2}$$

The stochastic error at iteration k of an algorithm based on the first-order information $\nabla f(x_k)$ is then $e_k = \nabla f(x_k) - \widehat{\nabla f}(x_k)$. Observe that conditioned on x_k, and by independent sampling, e_k has indeed zero-mean and variance that scales as $\mathcal{O}(1/m)$. Thus, to make this variance verify appropriate summability assumptions in k, that will be made clear in our analysis, one has to take m depend on k such that it increases fast enough.

1.1 Historical aspects

The Ravine method was introduced by Gelfand and Tsetlin [19] in 1961 in the case of a fixed positive extrapolation coefficient $\gamma_k \equiv \gamma > 0$. This method mimics the flow of water in the mountains which first flows rapidly downhill through small, steep ravines and then flows along the main river in the valley, hence its name. A geometric view of the Ravine Accelerated Gradient method is given in Figure 1.

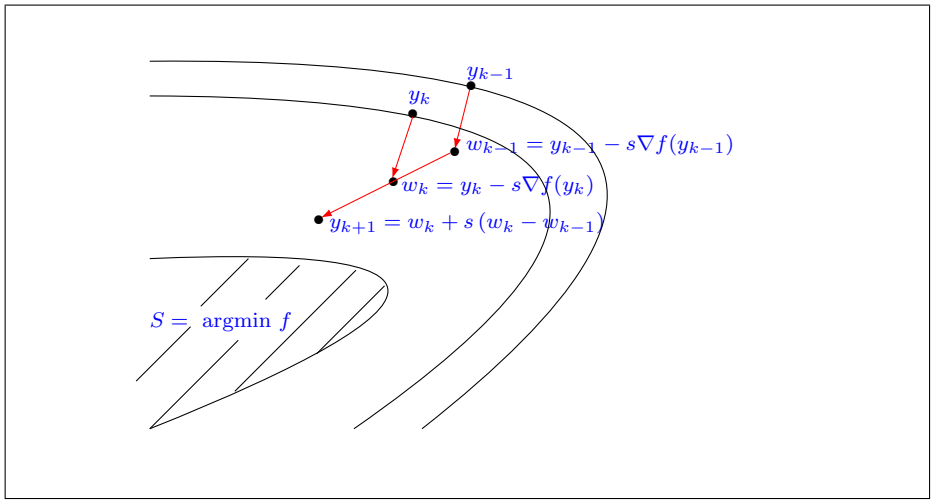

Fig. 1 (RAG): Ravine Accelerated Gradient method

The Ravine method was a precursor of the accelerated gradient methods. It has long been ignored but has recently appeared at the forefront of current research in numerical optimization, see for example Polyak [34], Attouch and Fadili [10], Shi, Du, Jordan and Su [38]. It comes naturally into the picture when considering the optimized first-order methods for smooth convex minimization, see [16,24,32].

When $\gamma_k = 1 - \frac{\alpha}{k}$, which, for $\alpha \geq 3$, the Ravine method is also closely related with the Nesterov accelerated gradient method [31,30], with which it has often been confused. In fact, the Ravine and Nesterov acceleration methods are both based on the operations of extrapolation and gradient descent, but in a reverse order. Furthermore, up to a slight change in the extrapolation coefficients, the two algorithms are associated with the same equations, each of them describing

the evolution of different variables, explaining how the two methods have been casually confused in some of the literature.

However, recent research concerning the understanding of accelerated first-order optimization methods, seen as temporal discretized dynamic systems, has made it possible to clarify the link between these two methods; see the recent work by some of the authors in [10,1]. In particular, these works have shown that while both algorithms share the same low-resolution ODE (i.e. of order 0 in the step-size), their super-resolution ODEs (i.e. of order 2 in the step-size) are fundamentally distinct. This was also confirmed by numerical experiments. This link will be further investigated for the discrete algorithms in Section 2.

1.2 Inertial stochastic gradient algorithms

Due to the importance of the subject in optimization, several works have been devoted to the study of perturbations in second-order dissipative inertial systems and in the corresponding first order algorithms (aka momentum methods). For deterministic perturbations, the subject was first considered for the case of a fixed viscous damping (aka heavy ball method with friction [33, 35]) in [9,22], then for the accelerated gradient method of Nesterov, and of the corresponding inertial dynamics with vanishing viscous damping, see [6,8,13,37,40].

Stochastic gradient descent methods with inertia are widely used in applications and at the core of optimization subroutines in many applications such as machine learning. Such algorithms are the subject of an active research work to understand their convergence behaviour and were studied in several works, focusing exclusively on stochastic versions of Nesterov's method and the heavy ball method; see [26,28,17,20,21,23,3,2,41,18,29,25,27,14,15].

1.3 Contributions

In this work, we propose and analyse both the stochastic Nesterov and Ravine methods with general extrapolation coefficient γ_k for solving infinite-dimensional optimization problems of the form $(\mathcal{P})$ in a real separable Hilbertian setting. In addition to the fact that this has not been done in the literature before–and in fact not for the Nesterov method as well, we are motivated by understanding the role of extrapolation on the convergence and stability properties of inertial systems. As we will explain in more detail later, not only taking a general coefficient γ_k gives a broad picture of the convergence properties of this class of algorithms, but also reveals the precise role of γ_k for balancing the trade-off between stability and fast convergence. Our contributions are the following:

- **Comprehensive convergence analysis for the Stochastic Ravine method with general extrapolation parameters**: we provide a unified analysis of the convergence properties of the Ravine method subject to noise in the gradient computation over a large class of the extrapolation sequence parameter. Previous analyses studied the setting of Nesterov's method where $\gamma_k = 1 - \frac{\alpha}{k}$. While we take inspiration from the work of [6] in the deterministic case, the extension from deterministic to stochastic errors requires a careful and comprehensive analysis.
- **Complexity estimates in expectation and almost sure sense**: we will establish fast convergence rates in expectation and in almost sure sense on the objective values (both in $\mathcal{O}(\cdot)$ and $o(\cdot)$) and on the gradient.
- **Weak convergence guarantees for the iterate sequence**: we will prove that the sequence of iterates provided by the Ravine method converges weakly almost surely to a random variable valued in the set of minimizers.

- **The discovery of what the influence of the extrapolation parameter is on the convergence properties.**: our results will highlight the trade-off between the decrease of the error variance and fast convergence of the values and gradients. In particular, some choices of the extrapolation parameter (and step-size sequences) will entail less stringent summability conditions on the error variance for convergence, but will result in slower a convergence rate, and vice-versa. We will see that a specific parametrization of the extrapolation parameters provides fast convergence properties of the Ravine algorithm resembling those of the Nesterov method. Moreover, our results show the flexibility of the method, the results being unchanged taking for example $\gamma_k = \frac{k}{k+\alpha}$ instead of $\gamma_k = 1 - \frac{\alpha}{k}$, as two of the many variations of the method.

1.4 Relation to prior work

We are not aware of any such a work for inertial algorithms (neither Nesterov nor Ravine) with general extrapolation coefficients applied to an infinite dimensional domain. Our complexity results are valid in expectation and almost surely. While the former is the standard in the literature, the latter is much less common, and the analysis less straightforward. Our results in expectation also cover some of those obtained by the works reviewed above (see the list in Section 1.2) as special cases when the extrapolation coefficients are those proposed by Nesterov (i.e., $\gamma_k = 1-\frac{\alpha}{k}$) and the heavy ball method (γ_k constant). In fact, even for these special cases, we complement the results of the literature with new ones. The almost sure weak convergence of the iterates is generally overlooked by most existing works (see Section 1.2) which focus exclusively on complexity estimates (except for the simple case of strongly convex objective functions).

1.5 A model result

Taking $\gamma_k = 1 - \frac{\alpha}{k}$ yields optimal convergence rate of the values and fast convergence of the gradients towards zero. Specifically, let the sequence $(y_k)_{k\in\mathbb{N}}$ generated by the stochastic Ravine method with constant step-size

$$\begin{cases} w_k = y_k - s(\nabla f(y_k) + e_k) \\ y_{k+1} = w_k + \left(1 - \frac{\alpha}{k}\right)(w_k - w_{k-1}), \end{cases}$$

where $s \in]0, 1/L]$, $(e_k)_{k\in\mathbb{N}}$ is a zero-mean stochastic noise. Let $\mathcal{F}_k$ be the sub-σ-algebra generated by y_0 and $(w_i)_{i\leq k-1}$. If $\alpha > 3$, $\mathbb{E}[e_k \mid \mathcal{F}_k] = 0$ and $\sum_{k=1}^{+\infty} k\mathbb{E}\left[\|e_k\|^2 \mid \mathcal{F}_k\right]^{1/2} < +\infty$ almost surely, then according to Theorem 3.2 and 3.3, the following convergence properties hold:

$$f(y_k) - \min_{\mathcal{H}} f = o\left(\frac{1}{k^2}\right) \quad \text{and} \quad \sum_k k^2 \|\nabla f(y_k)\|^2 < +\infty \quad \text{almost surely.}$$

In addition, if $\mathcal{H}$ is also separable[1], then the sequence $(y_k)_{k\in\mathbb{N}}$ converges weakly almost surely to a random variable valued in $\operatorname{argmin}(f)$. Our results in Section 4 will be established for a much larger lass of the extrapolation sequence beyond $1 - \alpha/k$. In particular, these results will emphasize the trade-off between the decrease of the error variance and fast convergence of the values and gradients.

[1] Separability is crucial for proving almost sure weak convergence of the sequence of iterates.

1.6 Contents

In Section 2, we start by making the link between the Ravine and the Nesterov method. This is instrumental because it makes it possible to transfer some known results of the Nesterov method. Section 3 is devoted to the study of the convergence properties of the stochastic Ravine method, with as an important result the fast convergence in mean of the gradients towards zero. Section 4 contains illustration and discussion of our results for various special choices of the extrapolation sequence γ_k. Finally we provide some conclusions.

2 Comparison of the Nesterov and Ravine methods

Let us first recall some basic facts concerning the Nesterov method.

2.1 Nesterov accelerated gradient method

The Nesterov Accelerated Gradient (NAG for short) method with general extrapolation coefficients $(\alpha_k)_{k\in\mathbb{N}}$, as studied in [6], reads

$$\begin{cases} y_k = x_k + \alpha_k(x_k - x_{k-1}) \\ x_{k+1} = y_k - s_k \nabla f(y_k). \end{cases} \qquad ((\mathrm{NAG})_{\alpha_k})$$

Its central role in optimization is due to the fact that a wise choice of the coefficients $(\alpha_k)_{k\in\mathbb{N}}$ provides an optimal convergence rate of the values (in the worst case).

Specifically, taking $\alpha_k = 1 - \frac{\alpha}{k}$ gives a scheme which, for $\alpha \geq 3$, generates iterates $(x_k)_{k\in\mathbb{N}}$ satisfying

$$f(x_k) - \min_{\mathcal{H}} f = \mathcal{O}\left(\frac{1}{k^2}\right) \text{ as } k \to +\infty, \tag{3}$$

and the fast convergence towards zero of the gradients (see [10])

$$\sum_k k^2 \|\nabla f(x_k)\|^2 < +\infty.$$

In addition, when $\alpha > 3$,

$$f(x_k) - \min_{\mathcal{H}} f = o\left(\frac{1}{k^2}\right) \text{ as } k \to +\infty, \ \sum_k k(f(x_k) - \min_{\mathcal{H}} f) < +\infty \tag{4}$$

and there is weak convergence of the iterates $(x_k)_{k\in\mathbb{N}}$ to optimal solutions, see [8,4,5,12,39].

2.2 Passing from Nesterov method to Ravine method and vice versa

To avoid confusion between the two algorithms $(\mathrm{RAG})_{\gamma_k}$ and $(\mathrm{NAG})_{\alpha_k}$, we use the subscript γ_k for the extrapolation coefficient in the Ravine method, and α_k for the extrapolation coefficient in the Nesterov method. A remarkable fact is that the variable y_k which enters the definition of $(\mathrm{NAG})_{\alpha_k}$ follows the $(\mathrm{RAG})_{\gamma_k}$ algorithm, with $\gamma_k = \alpha_{k+1}$. This generalizes the observation already made in [10] for the specific choice $\alpha_k = 1 - \frac{\alpha}{k}$. Although this is an elementary result, we give a detailed account of it in the following theorem, due to its importance.

Theorem 2.1 *(i) Let $(x_k)_{k\in\mathbb{N}}$ be the sequence generated by the Nesterov algorithm $(\mathrm{NAG})_{\alpha_k}$. Then the associated sequence $(y_k)_{k\in\mathbb{N}}$ also follows the equations of the Ravine algorithm $(\mathrm{RAG})_{\gamma_k}$ with $\gamma_k = \alpha_{k+1}$*

(ii) Conversely, if $(y_k)_{k\in\mathbb{N}}$ is the sequence associated to the Ravine method $(\mathrm{RAG})_{\gamma_k}$, then the sequence $(x_k)_{k\in\mathbb{N}}$ defined by $x_{k+1} := y_k - s_k\nabla f(y_k)$ follows the Nesterov algorithm $(\mathrm{NAG})_{\alpha_k}$ with $\alpha_k = \gamma_{k-1}$.

Proof (i) Suppose that $(x_k)_{k\in\mathbb{N}}$ follows $(\mathrm{NAG})_{\alpha_k}$. According to the definition of y_k

$$\begin{aligned} y_{k+1} &= x_{k+1} + \alpha_{k+1}(x_{k+1} - x_k) \\ &= y_k - s_k\nabla f(y_k) + \alpha_{k+1}\Big(y_k - s_k\nabla f(y_k) - (y_{k-1} - s_{k-1}\nabla f(y_{k-1}))\Big). \end{aligned}$$

Set $w_k := y_k - s_k\nabla f(y_k)$ (which is nothing but x_{k+1}). We obtain that $(y_k)_{k\in\mathbb{N}}$ follows $(\mathrm{RAG})_{\alpha_{k+1}}$, *i.e.*

$$(\mathrm{RAG})_{\alpha_{k+1}} \begin{cases} w_k = y_k - s_k\nabla f(y_k) \\ y_{k+1} = w_k + \alpha_{k+1}\,(w_k - w_{k-1})\,. \end{cases}$$

(ii) Conversely, suppose that $(y_k)_{k\in\mathbb{N}}$ follows the Ravine method $(\mathrm{RAG})_{\gamma_k}$. According to the definition of y_{k+1} and w_k, we have

$$y_{k+1} = y_k - s_k\nabla f(y_k) + \gamma_k\Big(y_k - s_k\nabla f(y_k) - (y_{k-1} - s_{k-1}\nabla f(y_{k-1}))\Big).$$

By definition of $x_{k+1} = y_k - s_k\nabla f(y_k)$, we deduce that

$$y_{k+1} = x_{k+1} + \gamma_k\,(x_{k+1} - x_k)\,.$$

Equivalently

$$y_k = x_k + \gamma_{k-1}\,(x_k - x_{k-1})\,.$$

Putting together the above relations and the definition of x_{k+1}, we obtain that $(x_k)_{k\in\mathbb{N}}$ follows $(\mathrm{NAG})_{\gamma_{k-1}}$, *i.e.*

$$(\mathrm{NAG})_{\gamma_{k-1}} \begin{cases} y_k = x_k + \gamma_{k-1}(x_k - x_{k-1}) \\ x_{k+1} = y_k - s_k\nabla f(y_k). \end{cases}$$

This completes the proof. □

Though the two methods are intimately linked as we have just seen, it is only recent advances in the dynamical system interpretation of the two methods that revealed their close relationship and also their differences. This is explained in the next section, where we consider the case of the Ravine method with general extrapolation coefficients, hence generalizing the work of [10] beyond the case $\alpha_k = 1 - \alpha/k$.

3 Convergence properties of the stochastic Ravine method

In this section, we analyze the convergence properties of the Ravine method with stochastic errors in the evaluation of the gradients. We first examine the fast convergence of the values and the convergence of iterates, then we show the fast convergence of the gradients towards zero. This section considers the algorithmic and stochastic version of the results obtained by the authors for the corresponding continuous dynamical systems with deterministic errors [11].

3.1 Values convergence rates and convergence of the iterates

We first start by proving the results for the Nesterov method before transferring them to the Ravine method thanks to Theorem 2.1. In [6], the Nesterov accelerated gradient method with a general extrapolation coefficient α_k and deterministic terms was studied. Here, we consider a stochastic version which reads for $k \geq 1$

$$\begin{cases} y_k = x_k + \alpha_k(x_k - x_{k-1}) \\ x_{k+1} = y_k - s_k(\nabla f(y_k) + e_k) \end{cases} \qquad ((\mathrm{SNAG})_{\alpha_k})$$

where $s_k \in]0, 1/L]$ is a sequence of step-sizes, $(e_k)_{k\in\mathbb{N}}$ is a sequence of $\mathcal{H}$-valued random variables. $(\mathrm{SNAG})_{\alpha_k}$ is initialized with $x_0 = x_1$, where x_0 a $\mathcal{H}$-valued, squared integrable random variable.

Taking the objective function $f \equiv 0$ and $e_k \equiv 0$ in $(\mathrm{SNAG})_{\alpha_k}$ already reveals insights for choosing the best parameters. In this case, the algorithm $(\mathrm{SNAG})_{\alpha_k}$ becomes $x_{k+1} - x_k - \alpha_k(x_k - x_{k-1}) = 0$. This implies that for every $k \geq 1$,

$$x_k = x_1 + \left(\sum_{i=1}^{k-1} \prod_{j=1}^{i} \alpha_j \right) (x_1 - x_0).$$

Therefore, $(x_k)_{k\in\mathbb{N}}$ converges if and only if $\sum_{i=1}^{+\infty} \prod_{j=1}^{i} \alpha_j < +\infty$. We are naturally led to introduce the sequence $(t_k)_{k\in\mathbb{N}}$ defined by

$$t_k := 1 + \sum_{i=k}^{+\infty} \prod_{j=k}^{i} \alpha_j. \tag{5}$$

The above formula may seem complicated at a first glance. In fact, the inverse transform, which makes it possible to pass from t_k to α_k has the following, simpler form

$$\alpha_k = \frac{t_k - 1}{t_{k+1}}. \tag{6}$$

Formula (6) will ease the path of the analysis and we shall make regular use of it in the sequel.

From now on, we denote by $(\Omega, \mathcal{F}, \mathbb{P})$ a probability space. We assume that $\mathcal{H}$ is a real separable Hilbert space endowed with its Borel σ-algebra, $\mathcal{B}(\mathcal{H})$. We denote a filtration on $(\Omega, \mathcal{F}, \mathbb{P})$ by $\mathscr{F} := (\mathcal{F}_k)_{k\in\mathbb{N}}$ where $\mathcal{F}_k$ is a sub-σ-algebra satisfying, for each $k \in \mathbb{N}$, $\mathcal{F}_k \subset \mathcal{F}_{k+1} \subset \mathcal{F}$. Furthermore, given a set of random variables $\{a_0, \ldots, a_k\}$ we denote by $\sigma(a_0, \ldots, a_k)$ the σ-algebra generated by $a_0, \ldots, a_k$. Finally, a statement (P) is said to hold ($\mathbb{P}$-a.s.) if

$$\mathbb{P}(\{\omega \in \Omega : (P) \text{ holds}\}) = 1.$$

Using the above notation, we denote the canonical filtration associated to the iterates of algorithm $(\mathrm{SNAG})_{\alpha_k}$ as $\mathscr{F}$ with, for all $k \in \mathbb{N}$,

$$\mathcal{F}_k := \sigma(x_0, \ldots, x_k)$$

such that all iterates up to x_k are completely determined by $\mathcal{F}_k$.

For the remainder of the paper, all equalities and inequalities involving random quantities should be understood as holding ($\mathbb{P}$-a.s.) even if it is not explicitly written.

Definition 3.1 Given a filtration $\mathscr{F}$, we denote by $\ell_+(\mathscr{F})$ the set of sequences of $[0,+\infty[$-valued random variables $(a_k)_{k\in\mathbb{N}}$ such that, for each $k\in\mathbb{N}$, a_k is $\mathcal{F}_k$-measurable. Then, for $p\in]0,+\infty[$, we also define the following set of p-summable random variables,

$$\ell_+^p(\mathscr{F}) := \left\{ (a_k)_{k\in\mathbb{N}} \in \ell_+(\mathscr{F}) : \sum_{k\in\mathbb{N}} a_k^p < +\infty \ \ (\mathbb{P}\text{-a.s.}) \right\}.$$

The set of non-negative p-summable (deterministic) sequences is denoted ℓ_+^p.

The following theorem is a generalization of [6, Theorems 3.1, 3.2 and 3.4] to the stochastic setting.

Theorem 3.1 *Assume that* (H) *holds and the sequence* $(\alpha_k)_{k\in\mathbb{N}}$ *satisfies*

$$\forall k\geq 1, \quad \sum_{i=k}^{+\infty}\prod_{j=k}^{i}\alpha_j < +\infty, \tag{K_0}$$

$$\forall k\geq 1, \quad t_{k+1}^2 - t_k^2 \leq t_{k+1}. \tag{K_1}$$

Consider the algorithm $(\mathrm{SNAG})_{\alpha_k}$ *where* $s_k\in]0,1/L]$ *is a non-increasing sequence and* $(e_k)_{k\in\mathbb{N}}$ *is a sequence of stochastic errors such that*

$$\mathbb{E}\left[e_k \mid \mathcal{F}_k\right] = 0 \ \ (\mathbb{P}\text{-}a.s.) \quad \text{and} \quad (s_k t_k \sigma_k)_{k\in\mathbb{N}} \in \ell_+^2(\mathscr{F}), \tag{K_2}$$

where $\sigma_k^2 := \mathbb{E}\left[\|e_k\|^2 \mid \mathcal{F}_k\right]$. *Then,*

(i) we have the following rate of convergence in almost sure and mean sense:

$$f(x_k) - \min f = \mathcal{O}\left(\frac{1}{s_k t_k^2}\right) \ (\mathbb{P}\text{-}a.s.),$$

and

$$\mathbb{E}\left[f(x_k)-\min f\right] \leq \frac{s_1 t_1^2 \mathbb{E}\left[f(x_0)-\min f\right] + \frac{1}{2}\mathbb{E}\left[\mathrm{dist}(x_0,S)^2\right] + 4\sum_{i=1}^{+\infty} s_i^2 t_i^2 \mathbb{E}\left[\|e_i\|^2\right]}{s_k t_k^2}.$$

(ii) Assume in addition that, for $m\in[0,1[$,

$$t_{k+1}^2 - t_k^2 \leq m\, t_{k+1} \quad \text{for every } k\geq 1, \tag{K_1^+}$$

then

$$\sum_{k\in\mathbb{N}} s_k t_{k+1}(f(x_k)-\min f) < +\infty \quad \text{and} \quad \sum_{k\in\mathbb{N}} t_k \|x_k - x_{k-1}\|^2 < +\infty \ \ (\mathbb{P}\text{-}a.s.).$$

If moreover $\sum_{k\in\mathbb{N}} \frac{t_{k+1}}{t_k^2} = +\infty$, *then*

$$f(x_k)-\min f = o\left(\frac{1}{s_k t_k^2}\right) \quad \text{and} \quad \|x_k - x_{k-1}\| = o\left(\frac{1}{t_k}\right) \ (\mathbb{P}\text{-}a.s.).$$

(iii) If $\alpha_k\in[0,1]$ *for every* $k\geq 1$, $\inf_k s_k > 0$, (K_1^+) *holds and* (K_2) *is strengthened to*

$$\mathbb{E}\left[e_k \mid \mathcal{F}_k\right] = 0 \ \ (\mathbb{P}\text{-}a.s.) \quad \text{and} \quad (s_k t_k \sigma_k)_{k\in\mathbb{N}} \in \ell_+^1(\mathscr{F}), \tag{K_2^+}$$

then the sequence $(x_k)_{k\in\mathbb{N}}$ *converges weakly* ($\mathbb{P}$-*a.s.*) *to an* $\mathrm{argmin}(f)$*-valued random variable.*

Before delving into the proof, some remarks are in order.

Remark 3.1 From claim (i), we have, for s_k constant and bounded away from 0, convergence at the rate $O(1/k^2)$ in the objective if $(t_k\sigma_k)_{k\in\mathbb{N}} \in \ell^2_+(\mathscr{F})$. If just $(\sigma_k)_{k\in\mathbb{N}} \in \ell^2_+(\mathscr{F})$, then the step-size must anneal at the rate $s_k \sim 1/t_k$ for an objective value convergence rate $O(1/t_k)$.

Now consider non-vanishing noise with bounded variance (i.e. $\limsup \sigma_k > 0$, $\mathbb{P}-a.s.$ and $\mathbb{E}[\sigma_k^2]-\mathbb{E}[\sigma_k]^2 \leq B$, $0 < B < \infty$). For the choice of $t_k = (k-1)/(\alpha-1)$, setting the step-size to be $s_k = 1/k^{1+\delta}$, with $\delta > 0$, results in convergence with a rate is $O(1/k^\delta)$. If $s_k = 1/k$ and the noise does not asymptotically vanish (a.s.), convergence can only be ensured to a noise dominated region. On the other hand, if $t_k = (k^{1+\delta}-1)/(\alpha-1)$ with $\delta < 0$, then $s_k = 1/k$ achieves a convergence rate of $O(1/k^\delta)$ if there is vanishing noise. Continuing, we see that the $O(1/k^2)$ rate is achieved for vanishing noise and $s_k = 1/k^{(2-\delta)}$.

The last statement of claim (ii) can be modified to get the same rate as in the deterministic case in [6, Theorem 3.4] but only at the price of a stronger summability assumption on the noise.

Proof Our proof is based on a (stochastic) Lyapunov analysis with appropriately chosen energy functionals.
(i) Denote $f_k(x) := f(x) + \langle e_k,\, x\rangle$ and recall $S = \operatorname{argmin}(f)$. Define the sequence

$$V_k := s_k t_k^2 (f(x_k) - f(x^\star)) + \frac{1}{2}\operatorname{dist}(z_k, S)^2 \text{ and } z_k := x_{k-1} + t_k\,(x_k - x_{k-1})\,.$$

Since f is convex and L-smooth, so is f_k. Let us apply (44) in Lemma A.3 on f_k successively at $y = y_k$ and $x = x_k$, then at $y = y_k$ and $x = x^\star \in S$. We get

$$f_k(x_{k+1}) \leq f_k(x_k) + \langle \nabla f_k(y_k),\, y_k - x_k\rangle - \frac{s_k}{2}\,\|\nabla f_k(y_k)\|^2 \tag{7}$$

$$f_k(x_{k+1}) \leq f_k(x^\star) + \langle \nabla f_k(y_k),\, y_k - x^\star\rangle - \frac{s_k}{2}\,\|\nabla f_k(y_k)\|^2\,. \tag{8}$$

Multiplying (7) by $t_{k+1}-1$ (which is non-negative by definition), then adding the (8), we derive that

$$\begin{aligned} t_{k+1} f_k(x_{k+1}) \leq (t_{k+1}-1) f_k(x_k) + f_k(x^\star) + \langle \nabla f_k(y_k),\, (t_{k+1}-1)(y_k - x_k) + y_k - x^\star\rangle \\ - \frac{s_k}{2} t_{k+1}\,\|\nabla f_k(y_k)\|^2\,. \end{aligned} \tag{9}$$

It is immediate to see, using (6) and the definitions of y_k and z_k, that

$$\begin{aligned} (t_{k+1}-1)(y_k - x_k) + y_k = x_k + t_{k+1}(y_k - x_k) &= x_{k-1} + (1 + t_{k+1}\alpha_k)(x_k - x_{k-1}) \\ &= x_{k-1} + t_k(x_k - x_{k-1}) = z_k. \end{aligned}$$

Inserting this into (9) and rearranging, we get

$$t_{k+1}(f_k(x_{k+1}) - f_k(x^\star)) \leq (t_{k+1}-1)(f_k(x_k) - f_k(x^\star)) + \langle \nabla f_k(y_k), z_k - x^\star\rangle - \frac{s_k}{2} t_{k+1}\,\|\nabla f_k(y_k)\|^2\,. \tag{10}$$

Straightforward computation, using again (6) and the definition of y_k and z_k, can yield the expression,

$$z_{k+1} - z_k = -s_k t_{k+1} \nabla f_k(y_k). \tag{11}$$

Thus

$$\|z_{k+1} - x^\star\|^2 = \|z_k - x^\star\|^2 - 2 s_k t_{k+1}\,\langle \nabla f_k(y_k),\, z_k - x^\star\rangle + s_k^2 t_{k+1}^2\,\|\nabla f_k(y_k)\|^2\,.$$

Dividing this by 2 and adding to (10), after multiplying the latter by $s_k t_{k+1}$, cancels all terms containing $\nabla f(y_k)$ and we arrive at

$$s_k t_{k+1}^2 (f_k(x_{k+1}) - f_k(x^\star)) + \frac{1}{2}\left\|z_{k+1} - x^\star\right\|^2 \leq s_k t_{k+1}(t_{k+1}-1)(f_k(x_k) - f_k(x^\star)) + \frac{1}{2}\left\|z_k - x^\star\right\|^2. \tag{12}$$

Let us take $x^\star$ as the closest point to z_k in S. Thus (12) is equivalent to

$$s_k t_{k+1}^2 (f_k(x_{k+1}) - f_k(x^\star)) + \frac{1}{2}\mathrm{dist}(z_{k+1}, S)^2 \leq s_k t_{k+1}(t_{k+1}-1)(f_k(x_k) - f_k(x^\star)) + \frac{1}{2}\mathrm{dist}(z_k, S)^2. \tag{13}$$

Let us now isolate the error terms. Inequality (13) is then equivalent to

$$\begin{aligned} s_k t_{k+1}^2 (f(x_{k+1}) - \min f) + \frac{1}{2}\mathrm{dist}(z_{k+1}, S)^2 \leq s_k t_{k+1}(t_{k+1}-1)(f(x_k) - \min f) + \frac{1}{2}\mathrm{dist}(z_k, S)^2 \\ - s_k \left\langle e_k,\, t_{k+1}^2 (x_{k+1} - x^\star) - t_{k+1}(t_{k+1}-1)(x_k - x^\star)\right\rangle. \end{aligned} \tag{14}$$

We have

$$t_{k+1}^2 (x_{k+1} - x^\star) - t_{k+1}(t_{k+1}-1)(x_k - x^\star) = t_{k+1}(z_{k+1} - x^\star).$$

In turn, using also that s_k is non-increasing, (14) becomes

$$\begin{aligned} s_{k+1} t_{k+1}^2 (f(x_{k+1}) - \min f) + \frac{1}{2}\mathrm{dist}(z_{k+1}, S)^2 + s_k(t_k^2 - t_{k+1}^2 + t_{k+1})(f(x_k) - \min f) \leq \\ s_k t_k^2 (f(x_k) - \min f) + \frac{1}{2}\mathrm{dist}(z_k, S)^2 - s_k t_{k+1} \left\langle e_k,\, z_{k+1} - x^\star\right\rangle. \end{aligned}$$

In view of the definition of V_k, this is equivalent to

$$V_{k+1} \leq V_k + s_k(t_{k+1}^2 - t_{k+1} - t_k^2)(f(x_k) - \min f) + s_k t_{k+1} \left\langle e_k,\, z_{k+1} - x^\star\right\rangle. \tag{15}$$

Taking the expectation conditionally on $\mathcal{F}_k$ in (15), we obtain

$$\mathbb{E}\left[V_{k+1} \mid \mathcal{F}_k\right] \leq V_k + s_k(t_{k+1}^2 - t_{k+1} - t_k^2)(f(x_k) - \min f) - s_k t_{k+1}\mathbb{E}\left[\left\langle e_k,\, z_{k+1} - x^\star\right\rangle \mid \mathcal{F}_k\right]. \tag{16}$$

We have

$$\begin{aligned} \mathbb{E}\left[\left\langle e_k,\, z_{k+1} - x^\star\right\rangle \mid \mathcal{F}_k\right] &= \mathbb{E}\left[\left\langle e_k,\, z_{k+1} - z_k\right\rangle \mid \mathcal{F}_k\right] + \left\langle \mathbb{E}\left[e_k \mid \mathcal{F}_k\right],\, z_k - x^\star\right\rangle \\ &= -s_k t_{k+1}\mathbb{E}\left[\left\langle e_k,\, \nabla f_k(y_k)\right\rangle \mid \mathcal{F}_k\right] = -s_k t_{k+1}\mathbb{E}\left[\left\langle e_k,\, \nabla f(y_k) + e_k\right\rangle \mid \mathcal{F}_k\right] \\ &= -s_k t_{k+1}\mathbb{E}\left[\left\|e_k\right\|^2 \mid \mathcal{F}_k\right] = -s_k t_{k+1}\sigma_k^2, \end{aligned}$$

where we used (11) in the second equality, and conditional unbiasedness (first part of (K_2)) in both the second and last inequalities, together with the fact that y_k, z_k and $x^\star$ are deterministic conditionally on $\mathcal{F}_k$. Plugging this into (16) yields

$$\begin{aligned} \mathbb{E}\left[V_{k+1} \mid \mathcal{F}_k\right] &\leq V_k + s_k(t_{k+1}^2 - t_{k+1} - t_k^2)(f(x_k) - \min f) + s_k^2 t_{k+1}^2 \sigma_k^2 \\ &\leq V_k + s_k(t_{k+1}^2 - t_{k+1} - t_k^2)(f(x_k) - \min f) + 4 s_k^2 t_k^2 \sigma_k^2, \end{aligned} \tag{17}$$

where we used that assumption (K_1) implies $t_{k+1} \leq 2t_k$; see [6, Remark 3.3]. Using again (K_1), the second term in the rhs of (17) is non-positive and can then be dropped. Now, thanks to the second part of (K_2), we are in position to apply Lemma A.1 to (17) to see that V_k converges

($\mathbb{P}$-a.s.), and consequently it is bounded ($\mathbb{P}$-a.s.). Thus, there exists a $[0,+\infty[$-valued random variable ξ such that $\sup_{k\in\mathbb{N}} V_k \leq \xi < +\infty$ ($\mathbb{P}$-a.s.). Therefore, for all $k \geq 1$,

$$s_k t_k^2 (f(x_k) - \min f) \leq V_k \leq \xi < +\infty \quad (\mathbb{P}\text{-a.s.}).$$

Moreover, taking the total expectation in (17) and iterating gives

$$s_k t_k^2 \mathbb{E}\left[f(x_k) - \min f\right] \leq \mathbb{E}\left[V_k\right] \leq \mathbb{E}\left[V_1\right] + 4\sum_{i=1}^{k} s_i^2 t_i^2 \mathbb{E}\left[\|e_i\|^2\right] \leq$$
$$s_1 t_1^2 \mathbb{E}\left[f(x_0) - \min f\right] + \frac{1}{2}\mathbb{E}\left[\operatorname{dist}(x_0, S)^2\right] + 4\sum_{i=1}^{+\infty} s_i^2 t_i^2 \mathbb{E}\left[\|e_i\|^2\right] < +\infty,$$

where we used in the last inequality that $x_0 = x_1$ by assumption, and that the rhs is finite thanks to Fubini-Tonelli's Theorem together with (K_2). This proves the first claim in the theorem.
(ii) Using (K_1^+) in (17), we get

$$\mathbb{E}\left[V_{k+1} \mid \mathcal{F}_k\right] \leq V_k - s_k(1-m)t_{k+1}(f(x_k) - \min f) + 4s_k^2 t_k^2 \sigma_k^2.$$

We can again invoke Lemma A.1 to get that

$$\sum_{k\geq 1} s_k t_{k+1}(f(x_k) - \min f) < +\infty \ (\mathbb{P}\text{-a.s.}). \tag{18}$$

Let

$$W_k := s_k(f(x_k) - \min f) + \frac{1}{2}\|x_k - x_{k-1}\|^2.$$

Combining [6, Proposition 2.1] with the fact that s_k is non-increasing, we have that

$$W_{k+1} \leq W_k - \frac{1-\alpha_k^2}{2}\|x_k - x_{k-1}\|^2 - s_k\langle e_k,\, x_{k+1} - x_k\rangle.$$

Taking the expectation conditionally on $\mathcal{F}_k$, we obtain

$$\mathbb{E}\left[W_{k+1} \mid \mathcal{F}_k\right] \leq W_k - \frac{1-\alpha_k^2}{2}\|x_k - x_{k-1}\|^2 - s_k\mathbb{E}\left[\langle e_k,\, x_{k+1} - x_k\rangle \mid \mathcal{F}_k\right]. \tag{19}$$

We have

$$\mathbb{E}\left[\langle e_k,\, x_{k+1} - x_k\rangle \mid \mathcal{F}_k\right] = \mathbb{E}\left[\langle e_k,\, x_{k+1} - y_k\rangle \mid \mathcal{F}_k\right] = -s_k\mathbb{E}\left[\langle e_k,\, \nabla f_k(y_k)\rangle \mid \mathcal{F}_k\right] = -s_k\mathbb{E}\left[\|e_k\|^2 \mid \mathcal{F}_k\right],$$

where we used the algorithm update of x_{k+1} in the second inequality, and conditional unbiasedness (first part of (K_2)) in the second and last inequalities together with x_k, y_k being conditionally deterministic on $\mathcal{F}_k$. Inserting this into (19) yields

$$\mathbb{E}\left[W_{k+1} \mid \mathcal{F}_k\right] \leq W_k - \frac{1-\alpha_k^2}{2}\|x_k - x_{k-1}\|^2 + s_k^2\sigma_k^2. \tag{20}$$

Multiplying (20) by t_{k+1}^2 and rearranging entails

$$\mathbb{E}\left[t_{k+1}^2 W_{k+1} \mid \mathcal{F}_k\right] \leq t_{k+1}^2 W_k - t_{k+1}^2\frac{1-\alpha_k^2}{2}\|x_k - x_{k-1}\|^2 + s_k^2 t_{k+1}^2\sigma_k^2$$
$$= t_k^2 W_k + s_k(t_{k+1}^2 - t_k^2)(f(x_k) - \min f) + \frac{t_{k+1}^2 - t_k^2 - t_{k+1}^2(1-\alpha_k^2)}{2}\|x_k - x_{k-1}\|^2 + s_k^2 t_{k+1}^2\sigma_k^2$$
$$\leq t_k^2 W_k + m s_k t_{k+1}(f(x_k) - \min f) - \frac{t_k}{2}\|x_k - x_{k-1}\|^2 + 4s_k^2 t_k^2\sigma_k^2. \tag{21}$$

In the equality, we used the expression of W_k. In the second inequality we used (K_1^+) and that $t_k = 1 + t_{k+1}\alpha_k$ and (K_1) which gives

$$t_{k+1}^2 - t_k^2 - t_{k+1}^2(1-\alpha_k^2) = (t_k-1)^2 - t_k^2 = -2t_k + 1 \leq -t_k$$

as $t_k \geq 1$. We have already proved above (see (18)) that $(s_k t_{k+1}(f(x_k) - \min f))_{k\in\mathbb{N}} \in \ell_+^1(\mathscr{F})$. Combining this with the second part of (K_2) allows us to invoke again Lemma A.1 on (21) to deduce that

$$\sum_{k\geq 1} t_k \left\|x_k - x_{k-1}\right\|^2 < +\infty \ (\mathbb{P}\text{-a.s.}). \tag{22}$$

Moreover, Lemma A.1 also implies that $t_k^2 W_k$ converges ($\mathbb{P}$-a.s.). On the other hand, we have

$$t_{k+1}W_k = s_k t_{k+1}(f(x_k) - \min f) + \frac{t_{k+1}}{2}\left\|x_k - x_{k-1}\right\|^2 \leq s_k t_{k+1}(f(x_k) - \min f) + t_k \left\|x_k - x_{k-1}\right\|^2,$$

and thus (18) and (22) imply that

$$\sum_{k\geq 1} t_{k+1} W_k < +\infty \ (\mathbb{P}\text{-a.s.}).$$

In turn

$$\sum_{k\geq 1} t_{k+1} W_k = \sum_{k\geq 1} \frac{t_{k+1}}{t_k^2} t_k^2 W_k < +\infty \quad (\mathbb{P}\text{-a.s.})$$

entailing that $\liminf_{k\to+\infty} t_k^2 W_k = 0$ ($\mathbb{P}$-a.s.). This together with ($\mathbb{P}$-a.s.) convergence of $t_k^2 W_k$ shown just above gives that

$$W_k = o\left(\frac{1}{t_k^2}\right).$$

Returning to the definition of W_k proves the assertions.

(iii) The crux of the proof consists in applying Opial's Lemma on a set of events of probability one. Observe that (K_2^+) implies (K_2). Thus Lemma A.1 applied to (21) ensures also that $t_k^2 W_k$ converges ($\mathbb{P}$-a.s.). In particular, this implies that $t_k \left\|x_k - x_{k-1}\right\|$ is bounded ($\mathbb{P}$-a.s.). From the proof of claim (i), we also know that ($\mathbb{P}$-a.s.), V_k converges, hence $(z_k)_{k\in\mathbb{N}}$ is bounded. In view of the definition of z_k, we obtain that $(x_k)_{k\in\mathbb{N}}$ is bounded ($\mathbb{P}$-a.s.). Moreover, since $t_k \geq 1$ and $\underline{s} = \inf_k s_k > 0$, we get from (ii) that ($\mathbb{P}$-a.s.)

$$\underline{s}\sum_{k\geq 1}(f(x_k) - \min f) \leq \sum_{k\geq 1} s_k t_{k+1}(f(x_k) - \min f) < +\infty,$$

and thus $\lim_{k\to+\infty} f(x_k) = \min f$ ($\mathbb{P}$-a.s.).

Let $\hat{\Omega}$ be the set of events on which the last statement holds and $\check{\Omega}$ on which boundedness of $(x_k)_{k\in\mathbb{N}}$ holds. Both sets are of probability one. For any $\omega \in \hat{\Omega} \cap \check{\Omega}$, let $(x_{k_j}(\omega))_{j\geq 1}$ be any converging subsequence, and $\bar{x}(\omega)$ its weak cluster point.

$$f(\bar{x}(\omega)) = \lim_{j\to\infty} f(x_{k_j}(\omega)) = \lim_{k\to\infty} f(x_k(\omega)) = \min f,$$

which means that $\bar{x}(\omega) \in S$. This implies that ($\mathbb{P}$-a.s.) each weak cluster point of $(x_k)_{k\in\mathbb{N}}$ belongs to $S = \operatorname{argmin}(f)$. In other words, the second condition of Opial's lemma holds ($\mathbb{P}$-a.s.).

Let $x^\star \in S$ and define $h_k := \frac{1}{2}\|x_k - x^\star\|^2$. We now show that $\lim_{k\to+\infty} h_k$ exists ($\mathbb{P}$-a.s.). For this, we use a standard argument that can be found e.g. in [12,6]. By [6, Proposition 2.3], we have

$$\begin{aligned} h_{k+1} - h_k - \alpha_k(h_k - h_{k-1}) &\leq \frac{\alpha_k(1+\alpha_k)}{2}\|x_k - x_{k-1}\|^2 - s_k(f_k(x_{k+1}) - f_k(x^\star) \\ &\leq \|x_k - x_{k-1}\|^2 - s_k(f(x_{k+1}) - \min f) - s_k\langle e_k,\, x_{k+1} - x^\star\rangle \\ &\leq \|x_k - x_{k-1}\|^2 - s_k\langle e_k,\, x_{k+1} - x^\star\rangle\,. \end{aligned}$$

In the second inequality we used that $\alpha_k \in [0,1]$, and the last one minimality of $x^\star$. Almost sure boundedness of x_k implies that there exists a $[0,+\infty[$-valued random variable η such that $\sup_{k\in\mathbb{N}}\|x_k - x^\star\| \leq \eta < +\infty$ ($\mathbb{P}$-a.s.). Thus

$$h_{k+1} - h_k - \alpha_k(h_k - h_{k-1}) \leq \|x_k - x_{k-1}\|^2 + \eta s_k\|e_k\|\,. \tag{23}$$

Multiplying (23) by t_{k+1}, taking the positive part and the conditional expectation, we end up having

$$\begin{aligned} \mathbb{E}\left[t_{k+1}(h_{k+1} - h_k)_+ \mid \mathcal{F}_k\right] &\leq t_{k+1}\alpha_k(h_k - h_{k-1})_+ + t_{k+1}\|x_k - x_{k-1}\|^2 + \eta s_k t_{k+1}\mathbb{E}\left[\|e_k\| \mid \mathcal{F}_k\right] \\ &\leq (t_k - 1)(h_k - h_{k-1})_+ + t_{k+1}\|x_k - x_{k-1}\|^2 + 2\eta s_k t_k \mathbb{E}\left[\|e_k\|^2 \mid \mathcal{F}_k\right]^{1/2} \\ &= t_k(h_k - h_{k-1})_+ - (h_k - h_{k-1})_+ + t_{k+1}\|x_k - x_{k-1}\|^2 + 2\eta s_k t_k \sigma_k. \end{aligned}$$

where we used that $t_k = 1 + t_{k+1}\alpha_k$, that $t_{k+1} \leq 2t_k$ and Jensen's inequality. As the last two terms in the rhs are summable ($\mathbb{P}$-a.s.), we get using Lemma A.1 that $((h_k - h_{k-1})_+)_{k\in\mathbb{N}} \in \ell^1_+(\mathscr{F})$ ($\mathbb{P}$-a.s.). In turn, since h_k is non-negative, we get by a classical argument that $\lim_{k\to+\infty} h_k$ exists.

Note that the set of events of probability on which $\lim_{k\to+\infty} h_k$ exists depends on $x^\star$. To make this uniform on S we use a separability argument.

Indeed, we have just shown that there exists a set of events $\Omega_{x^\star}$ (that depends on $x^\star$) such that $\mathbb{P}(\Omega_{x^\star}) = 1$ and for all $\omega \in \Omega_{x^\star}$, $(\|x_k(\omega) - x^\star\|)_{k\in\mathbb{N}}$ converges. We now show that there exists a set of events independent of $x^\star$, whose probability is one and such that the above still holds on this set. Since $\mathcal{H}$ is separable, there exists a countable set $U \subseteq S$, such that $\mathrm{cl}(U) = S$. Let $\tilde{\Omega} = \bigcap_{u\in U}\Omega_u$. Since U is countable, a union bound shows

$$\mathbb{P}(\tilde{\Omega}) = 1 - \mathbb{P}\left(\bigcup_{u\in U}\Omega_u^c\right) \geq 1 - \sum_{u\in U}\mathbb{P}(\Omega_u^c) = 1.$$

For arbitrary $x^\star \in S$, there exists a sequence $(u_j)_{j\in\mathbb{N}} \subset U$ such that u_j converges strongly to $x^\star$. Thus for every $j \in \mathbb{N}$ there exists $\tau_j : \Omega_{u_j} \to \mathbb{R}_+$ such that

$$\lim_{k\to+\infty}\|x_k(\omega) - u_j\| = \tau_j(\omega), \quad \forall\omega \in \Omega_{u_j}. \tag{24}$$

Now, let $\omega \in \tilde{\Omega}$. Since $\tilde{\Omega} \subset \Omega_{u_j}$ for any $j \geq 1$, and using the triangle inequality and (24), we obtain that

$$\tau_j(\omega) - \|u_j - x^\star\| \leq \liminf_{k\to+\infty}\|x_k(\omega) - x^\star\| \leq \limsup_{k\to+\infty}\|x_k(\omega) - x^\star\| \leq \tau_j(\omega) + \|u_j - x^\star\|\,.$$

Passing to $j \to +\infty$, we deduce

$$\limsup_{j\to+\infty}\tau_j(\omega) \leq \liminf_{k\to+\infty}\|x_k(\omega) - x^\star\| \leq \limsup_{k\to+\infty}\|x_k(\omega) - x^\star\| \leq \liminf_{j\to+\infty}\tau_j(\omega),$$

whence we deduce that $\lim_{j\to+\infty}\tau_j(\omega)$ exists for all $\omega\in\tilde{\Omega}$. In turn, ($\mathbb{P}$-a.s.), $\lim_{k\to+\infty}\|x_k-x^\star\|$ exists and is equal to $\lim_{j\to+\infty}\tau_j$ for any $x^\star\in S$.

We are now in position to apply Opial's Lemma at any $\omega\in\hat{\Omega}\cap\check{\Omega}\cap\tilde{\Omega}$, since $\mathbb{P}(\hat{\Omega}\cap\check{\Omega}\cap\tilde{\Omega})=1$, to conclude. □

Let us now return to the Ravine algorithm. A simple adaptation of the proof of Theorem 2.1 applied to $(\mathrm{SNAG})_{\alpha_k}$ (just replace f by $f+\langle e_k,\cdot\rangle$, and follow similar algebraic manipulations) gives that the associated sequence $(y_k)_{k\in\mathbb{N}}$ defined by

$$y_k = x_k + \alpha_k(x_k - x_{k-1}),$$

follows the stochastic Ravine accelerated gradient algorithm with $\gamma_k=\alpha_{k+1}$, i.e. for all $k\geq 1$

$$\begin{cases} w_k = y_k - s_k(\nabla f(y_k)+e_k) \\ y_{k+1} = w_k + \alpha_{k+1}\left(w_k - w_{k-1}\right). \end{cases} \qquad ((\mathrm{SRAG})_{\alpha_{k+1}})$$

$(\mathrm{SRAG})_{\alpha_{k+1}}$ is initialized with y_0 and $w_{-1}=y_0$, where y_0 is a $\mathcal{H}$-valued, squared integrable random variable. According to this relationship between the Nesterov and the Ravine method highlighted in Theorem 2.1, the results of Theorem 3.1 can now be transposed to $(\mathrm{SRAG})_{\alpha_{k+1}}$. For this, we denote the canonical filtration associated to $(\mathrm{SRAG})_{\alpha_{k+1}}$ as $\mathscr{F}=(\mathcal{F}_k)_{k\in\mathbb{N}}$ with, $\forall k\geq\mathbb{N}$, $\mathcal{F}_k=\sigma(y_0,(w_i)_{i\leq k-1})$.

Theorem 3.2 *Assume the conditions presented in* (H). *Let $(y_k)_{k\in\mathbb{N}}$ be the sequence generated by $(\mathrm{SRAG})_{\alpha_{k+1}}$ where $s_k\in]0,1/L]$ is a non-increasing sequence, $(\alpha_k)_{k\in\mathbb{N}}\subset[0,1]$ satisfies (K_0) and (K_1^+) with $\sum_{k\in\mathbb{N}}\frac{t_{k+1}}{t_k^2}=+\infty$, and $(e_k)_{k\in\mathbb{N}}$ is a sequence of stochastic errors satisfying (K_2^+). Then, the sequence $(y_k)_{k\in\mathbb{N}}$ satisfies*

$$\sum_{k\in\mathbb{N}} s_k t_{k+1}(f(y_k)-\min f)<+\infty \quad \textit{and} \quad f(y_k)-\min_{\mathcal{H}} f = o\left(\frac{1}{s_k t_k^2}\right) \quad \textit{as } k\to+\infty \quad (\mathbb{P}\textit{-a.s.}).$$

Moreover, if $\inf_k s_k>0$, then the sequence $(y_k)_{k\in\mathbb{N}}$ converges weakly ($\mathbb{P}$-a.s.) to an $\operatorname{argmin}(f)$-valued random variable.

Proof According to Theorem 2.1, the sequence $(x_k)_{k\in\mathbb{N}}$ defined by

$$x_{k+1} = y_k - s_k(\nabla f(y_k)+e_k) \tag{25}$$

is equivalent to Algorithm $(\mathrm{SNAG})_{\alpha_k}$. It then follows from Theorem 3.1(ii) that

$$f(x_k)-\min f = o\left(\frac{1}{s_k t_k^2}\right) \quad \text{and} \quad \|x_k-x_{k-1}\| = o\left(\frac{1}{t_k}\right) \quad (\mathbb{P}\text{-a.s.}). \tag{26}$$

In addition, in view of condition (K_2^+), we can apply Lemma A.2 with $\varepsilon_k=(s_k t_k\sigma_k)_{k\in\mathbb{N}}$ to infer that

$$\sum_{k=1}^{+\infty} s_k t_k \|e_k\| < +\infty \quad (\mathbb{P}\text{-a.s.}), \tag{27}$$

and thus

$$s_k\|e_k\| = o\left(\frac{1}{t_k}\right) \quad (\mathbb{P}\text{-a.s.}). \tag{28}$$

Rearrange the terms in (25) to obtain the expression $\nabla f(y_k) = -\frac{1}{s_k}(x_{k+1} - y_k) - e_k$. Using, successively, the convexity of f, the Cauchy-Schwartz inequality, and the triangle inequality, we obtain

$$\begin{aligned} f(y_k) - \min_{\mathcal{H}} f &\leq f(x_k) - \min_{\mathcal{H}} f + \frac{1}{s_k} \langle x_{k+1} - y_k + s_k e_k, x_k - y_k \rangle \\ &\leq f(x_k) - \min_{\mathcal{H}} f + \frac{1}{s_k} \left(\|x_{k+1} - y_k\| + s_k \|e_k\| \right) \|x_k - y_k\| \\ &\leq f(x_k) - \min_{\mathcal{H}} f + \frac{1}{s_k} \left(\|x_{k+1} - x_k\| + \|x_k - y_k\| + s_k \|e_k\| \right) \|x_k - y_k\|. \quad (29) \end{aligned}$$

Using again the link between $(\mathrm{SRAG})_{\alpha_{k+1}}$ and $(\mathrm{SNAG})_{\alpha_k}$, we have

$$y_k = x_k + \alpha_k (x_k - x_{k-1}).$$

Therefore, since $\alpha_k \in [0, 1]$,

$$\|y_k - x_k\| \leq \|x_k - x_{k-1}\|. \quad (30)$$

Combining (26), (28), (29) and (30) we obtain

$$\begin{aligned} f(y_k) - \min_{\mathcal{H}} f &\leq f(x_k) - \min_{\mathcal{H}} f + \frac{1}{s_k} \left(\|x_{k+1} - x_k\| + \|x_k - x_{k-1}\| + s_k \|e_k\| \right) \|x_k - x_{k-1}\| \\ &= o\left(\frac{1}{s_k t_k^2} \right) \quad (\mathbb{P}\text{-a.s.}) \end{aligned}$$

where we used that $t_{k+1} \leq 2t_k$ in the last equality. In addition, using Young's inequality, that $(x_k)_{k\in\mathbb{N}}$ is bounded ($\mathbb{P}$-a.s.), (27) and the summability claims of Theorem 3.1(ii), we get that ($\mathbb{P}$-a.s.),

$$\begin{aligned} \sum_{k\in\mathbb{N}} s_k t_{k+1} (f(y_k) - \min f) \leq \sum_{k\in\mathbb{N}} s_k t_{k+1} (f(x_k) - \min f) + \sum_{k\in\mathbb{N}} \frac{t_{k+1}}{2} \|x_{k+1} - x_k\|^2 \\ + 3 \sum_{k\in\mathbb{N}} t_k \|x_k - x_{k-1}\|^2 + 4\eta \sum_{k\in\mathbb{N}} t_k s_k \|e_k\| < +\infty, \end{aligned}$$

where η is the $[0, +\infty[$-valued random variable such that $\sup_{k\in\mathbb{N}} \|x_k\| \leq \eta < +\infty$ ($\mathbb{P}$-a.s.).

Now, from (26) and (30), we also have $\|y_k - x_k\| = o\left(\frac{1}{t_k}\right)$ ($\mathbb{P}$-a.s.). Consequently, $y_k - x_k$ converges strongly ($\mathbb{P}$-a.s.) to zero. Since the sequence $(x_k)_{k\in\mathbb{N}}$ converges weakly, it follows that the sequence $(y_k)_{k\in\mathbb{N}}$ converges weakly ($\mathbb{P}$-a.s.) to the same limit as $(x_k)_{k\in\mathbb{N}}$, and we know from Theorem 3.1(iii) that the latter indeed converges weakly ($\mathbb{P}$-a.s.) to an argmin(f)-valued random variable. □

3.2 Fast convergence of the gradients towards zero

In this section, the previous results on the stochastic Ravine method $(\mathrm{SRAG})_{\alpha_{k+1}}$ are completed in also showing the fast convergence towards zero of the gradients. This will necessitate a specific and intricate Lyapunov analysis[2].

[2] Observe that embarking from (7)-(8) and using the refined estimate in (44) is not sufficient to get the result.

Recall $f_k(x) := f(x) + \langle e_k, x\rangle$ from the proof of Theorem 3.1. The formula in Lemma 3.1 hereafter will play a key role in our Lyapunov analysis, and will serve as the constitutive formulation of the algorithm. It corresponds to the Hamiltonian formulation of the algorithm involving the discrete velocities which are defined by, for each $k \in \mathbb{N}$

$$v_k := \frac{1}{h}(y_k - y_{k-1}) \tag{31}$$

where we recall that $h = \sqrt{s}$.

Lemma 3.1 *Let $(y_k)_{k\in\mathbb{N}}$ be generated by $(\mathrm{SRAG})_{\alpha_{k+1}}$. Then, for all $k \in \mathbb{N}$*

$$t_{k+1}(v_k + h\nabla f_{k-1}(y_{k-1})) - (t_k - 1)(v_{k-1} + h\nabla f_{k-2}(y_{k-2})) = -h(t_k - 1)\nabla f_{k-1}(y_{k-1}). \tag{32}$$

Proof According to the algorithm recursion, we have

$$\begin{aligned} y_k &= y_{k-1} - h^2\nabla f_{k-1}(y_{k-1}) + \alpha_k\left(y_{k-1} - h^2\nabla f_{k-1}(y_{k-1}) - \left(y_{k-2} - h^2\nabla f_{k-2}(y_{k-2})\right)\right) \\ &= y_{k-1} + \alpha_k(y_{k-1} - y_{k-2}) - h^2\left(\nabla f_{k-1}(y_{k-1}) + \alpha_k\left(\nabla f_{k-1}(y_{k-1}) - \nabla f_{k-2}(y_{k-2})\right)\right). \end{aligned}$$

Equivalently,

$$\begin{aligned} 0 &= (y_k - y_{k-1}) - \alpha_k(y_{k-1} - y_{k-2}) + h^2\nabla f_{k-1}(y_{k-1}) + h^2\alpha_k(\nabla f_{k-1}(y_{k-1}) - \nabla f_{k-2}(y_{k-2})) \\ &= \alpha_k(y_k - y_{k-1}) - \alpha_k(y_{k-1} - y_{k-2}) + (1-\alpha_k)(y_k - y_{k-1}) + h^2\nabla f_{k-1}(y_{k-1}) \\ &\quad + h^2\alpha_k(\nabla f_{k-1}(y_{k-1}) - \nabla f_{k-2}(y_{k-2})). \end{aligned}$$

Let us make v_k appear by multiplying this equality by $\frac{1}{h\alpha_k}$. We then get

$$\begin{aligned} 0 &= v_k - v_{k-1} + \frac{1-\alpha_k}{\alpha_k}v_k + \frac{h}{\alpha_k}\nabla f_{k-1}(y_{k-1}) + h(\nabla f_{k-1}(y_{k-1}) - \nabla f_{k-2}(y_{k-2})) \\ &= (v_k + h\nabla f_{k-1}(y_{k-1})) - (v_{k-1} + h\nabla f_{k-2}(y_{k-2})) + \frac{1-\alpha_k}{\alpha_k}v_k + \frac{h}{\alpha_k}\nabla f_{k-1}(y_{k-1}). \end{aligned}$$

After multiplication by $\frac{\alpha_k}{1-\alpha_k}$, we arrive at

$$\begin{aligned} 0 &= \frac{\alpha_k}{1-\alpha_k}(v_k + h\nabla f_{k-1}(y_{k-1})) - \frac{\alpha_k}{1-\alpha_k}(v_{k-1} + h\nabla f_{k-2}(y_{k-2})) + v_k + \frac{h}{1-\alpha_k}\nabla f_{k-1}(y_{k-1}) \\ &= \left(1 + \frac{\alpha_k}{1-\alpha_k}\right)(v_k + h\nabla f_{k-1}(y_{k-1})) - \frac{\alpha_k}{1-\alpha_k}(v_{k-1} + h\nabla f_{k-2}(y_{k-2})) - h\nabla f_{k-1}(y_{k-1}) \\ &\qquad + \frac{h}{1-\alpha_k}\nabla f_{k-1}(y_{k-1}). \end{aligned}$$

We thus obtain

$$\frac{1}{1-\alpha_k}(v_k + h\nabla f_{k-1}(y_{k-1})) - \frac{\alpha_k}{1-\alpha_k}(v_{k-1} + h\nabla f_{k-2}(y_{k-2})) = -\frac{h\alpha_k}{1-\alpha_k}\nabla f_{k-1}(y_{k-1}).$$

Equivalently

$$(v_k + h\nabla f_{k-1}(y_{k-1})) - \alpha_k(v_{k-1} + h\nabla f_{k-2}(y_{k-2})) = -h\alpha_k\nabla f_{k-1}(y_{k-1}). \tag{33}$$

In view of (6), the last equality is also equivalent to (32). This completes the proof of the Lemma. □

Recall the canonical filtration associated to $(\mathrm{SRAG})_{\alpha_{k+1}}$ as $\mathscr{F} = (\mathcal{F}_k)_{k\in\mathbb{N}}$ with, $\forall k \geq \mathbb{N}$, $\mathcal{F}_k = \sigma(y_0, (w_i)_{i\leq k-1})$.

Theorem 3.3 *Let us assume the conditions defined in* (H). *Let* $(y_k)_{k\in\mathbb{N}}$ *be the sequence generated by* $(\mathrm{SRAG})_{\alpha_{k+1}}$ *where* $s_k \equiv s \in]0, 1/L]$, $(\alpha_k)_{k\in\mathbb{N}} \subset [0,1]$ *satisfy* (K_0) *and* (K_1^+). *Assume that* $(e_k)_{k\in\mathbb{N}}$ *is a sequence of stochastic errors subject to conditions* (K_2^+). *Then the sequence of gradients* $(\nabla f(y_k))_{k\in\mathbb{N}}$ *converges to zero with*

$$\sum_{k\in\mathbb{N}} t_{k+1}^2 \|\nabla f(y_k)\|^2 < +\infty \quad (\mathbb{P}\text{-}a.s.)\,.$$

Proof Our Lyapunov analysis is based on the sequence $(E_k)_{k\in\mathbb{N}}$ defined as

$$E_k := h^2(t_{k+1}-1)t_{k+1}(f(y_{k-1}) - \min f) + \frac{1}{2}\mathrm{dist}(z_k, S)^2,$$
$$z_k := y_k + h(t_{k+1}-1)\Big(v_k + h\nabla f_{k-1}(y_{k-1})\Big).$$

Let $x^\star$ be the closest point to z_k in S. By definition of E_k, we have

$$\begin{aligned} &E_{k+1} - E_k \leq h^2(t_{k+1}-1)t_{k+1}(f(y_k) - f(y_{k-1})) \\ &+ h^2\Big((t_{k+2}-1)t_{k+2} - (t_{k+1}-1)t_{k+1}\Big)(f(y_k) - \min f) + \frac{1}{2}\|z_{k+1} - x^\star\|^2 - \frac{1}{2}\|z_k - x^\star\|^2. \end{aligned} \tag{34}$$

Let us compute this last expression with the help of the elementary inequality

$$\frac{1}{2}\|z_{k+1} - x^\star\|^2 - \frac{1}{2}\|z_k - x^\star\|^2 = \langle z_{k+1} - z_k, z_{k+1} - x^\star\rangle - \frac{1}{2}\|z_{k+1} - z_k\|^2. \tag{35}$$

Recall the constitutive equation given by (32) that we write as follows

$$t_{k+2}(v_{k+1} + h\nabla f_k(y_k)) - (t_{k+1}-1)(v_k + h\nabla f_{k-1}(y_{k-1})) = -h(t_{k+1}-1)\nabla f_k(y_k). \tag{36}$$

Using successively the definition of z_k and (36), we obtain

$$\begin{aligned} z_{k+1} - z_k &= (y_{k+1} - y_k) + h(t_{k+2}-1)\Big(v_{k+1} + h\nabla f_k(y_k)\Big) - h(t_{k+1}-1)\Big(v_k + h\nabla f_{k-1}(y_{k-1})\Big) \\ &= hv_{k+1} - h\Big(v_{k+1} + h\nabla f_k(y_k)\Big) - h^2(t_{k+1}-1)\nabla f_k(y_k) = -h^2 t_{k+1}\nabla f_k(y_k). \end{aligned}$$

This together with the definition of z_k yields

$$z_{k+1} = z_k - h^2 t_{k+1}\nabla f_k(y_k) = y_k + h(t_{k+1}-1)\Big(v_k + h\nabla f_{k-1}(y_{k-1})\Big) - h^2 t_{k+1}\nabla f_k(y_k).$$

Plugging this into (35), we deduce that

$$\begin{aligned} &\frac{1}{2}\|z_{k+1} - x^\star\|^2 - \frac{1}{2}\|z_k - x^\star\|^2 = -\frac{1}{2}h^4 t_{k+1}^2\|\nabla f_k(y_k)\|^2 \\ &-h^2 t_{k+1}\Big\langle \nabla f_k(y_k), y_k - x^\star + h(t_{k+1}-1)\Big(v_k + h\nabla f_{k-1}(y_{k-1})\Big) - h^2 t_{k+1}\nabla f_k(y_k)\Big\rangle \\ &= \frac{1}{2}h^4 t_{k+1}^2\|\nabla f_k(y_k)\|^2 - h^2 t_{k+1}\Big\langle \nabla f_k(y_k), y_k - x^\star + h(t_{k+1}-1)\Big(v_k + h\nabla f_{k-1}(y_{k-1})\Big)\Big\rangle. \end{aligned}$$

Let us arrange the above expression so as to group the products of $\nabla f_k(y_k)$. For this, we use (32) again, written as,

$$(t_{k+1}-1)(v_k + h\nabla f_{k-1}(y_{k-1})) = t_{k+2}(v_{k+1} + h\nabla f_k(y_k)) + h(t_{k+1}-1)\nabla f_k(y_k). \tag{37}$$

Therefore,

$$\begin{aligned}
&y_k - x^\star + h(t_{k+1}-1)\Big(v_k + h\nabla f_{k-1}(y_{k-1})\Big)\\
&= y_k - x^\star + ht_{k+2}(v_{k+1} + h\nabla f_k(y_k)) + h^2(t_{k+1}-1)\nabla f_k(y_k)\\
&= y_k - x^\star + ht_{k+2}v_{k+1} + h^2(t_{k+2}+t_{k+1}-1)\nabla f_k(y_k).
\end{aligned}$$

Collecting the above results we obtain

$$\begin{aligned}
\frac{1}{2}\|z_{k+1}-x^\star\|^2 - \frac{1}{2}\|z_k - x^\star\|^2 &= \frac{1}{2}h^4 t_{k+1}^2\|\nabla f_k(y_k)\|^2\\
&\quad - h^2 t_{k+1}\left\langle \nabla f_k(y_k), y_k - x^\star + ht_{k+2}v_{k+1} + h^2(t_{k+2}+t_{k+1}-1)\nabla f_k(y_k)\right\rangle.
\end{aligned}$$

Inserting this in (34) we get

$$\begin{aligned}
&E_{k+1} - E_k \le h^2(t_{k+1}-1)t_{k+1}(f(y_k) - f(y_{k-1}))\\
&+ h^2\Big((t_{k+2}-1)t_{k+2} - (t_{k+1}-1)t_{k+1}\Big)(f(y_k) - \min f) + \frac{1}{2}h^4 t_{k+1}^2\|\nabla f_k(y_k)\|^2\\
&- h^2 t_{k+1}\left\langle \nabla f_k(y_k), y_k - x^\star + ht_{k+2}v_{k+1} + h^2(t_{k+2}+t_{k+1}-1)\nabla f_k(y_k)\right\rangle. \qquad (38)
\end{aligned}$$

In view of the basic gradient inequality for convex differentiable functions whose gradient is L-Lipschitz continuous, we have

$$\begin{aligned}
f(y_{k-1}) &\ge f(y_k) + \langle \nabla f(y_k), y_{k-1} - y_k\rangle + \frac{1}{2L}\|\nabla f(y_k) - \nabla f(y_{k-1})\|^2.\\
\min f &\ge f(y_k) + \langle \nabla f(y_k), x^\star - y_k\rangle + \frac{1}{2L}\|\nabla f(y_k)\|^2.
\end{aligned}$$

Combining the above inequalities with (38), and using $\nabla f_k(y_k) = \nabla f(y_k) + e_k$, we get

$$\begin{aligned}
&E_{k+1} - E_k \le -h^2(t_{k+1}-1)t_{k+1}\Big(\langle \nabla f(y_k), y_{k-1} - y_k\rangle + \frac{1}{2L}\|\nabla f(y_k) - \nabla f(y_{k-1})\|^2\Big)\\
&+ h^2\Big((t_{k+2}-1)t_{k+2} - (t_{k+1}-1)t_{k+1}\Big)(f(y_k) - \min f) - h^2 t_{k+1}(f(y_k) - \min f)\\
&+ \frac{1}{2}h^4 t_{k+1}^2\|\nabla f_k(y_k)\|^2 - h^2 t_{k+1}\left\langle \nabla f_k(y_k), ht_{k+2}v_{k+1} + h^2(t_{k+2}+t_{k+1}-1)\nabla f_k(y_k)\right\rangle\\
&- h^2 t_{k+1}\langle y_k - x^\star, e_k\rangle. \qquad (39)
\end{aligned}$$

Next rearrange the last inequality by grouping terms on the right hand side with common expressions. To begin with, rewrite the second and third summand as follows:

$$\begin{aligned}
h^2\Big((t_{k+2}-1)t_{k+2} - (t_{k+1}-1)t_{k+1}\Big)(f(y_k) - \min f) - h^2 t_{k+1}(f(y_k) - \min f) &=\\
- h^2\Big(t_{k+1}^2 - t_{k+2}^2 + t_{k+2}\Big)(f(y_k) - \min f).&
\end{aligned}$$

For the following expression grouping two of the summands above, we use the definition of v_k for the first equality, and the constitutive equation (37) for the third,

$$\begin{aligned}
&- h^2(t_{k+1}-1)t_{k+1}\langle \nabla f(y_k), y_{k-1} - y_k\rangle - h^2 t_{k+1}\langle \nabla f_k(y_k), ht_{k+2}v_{k+1}\rangle\\
&= h^3(t_{k+1}-1)t_{k+1}\langle \nabla f(y_k), v_k\rangle - h^3 t_{k+1}t_{k+2}\langle \nabla f_k(y_k), v_{k+1}\rangle\\
&= h^3 t_{k+1}\langle \nabla f(y_k), (t_{k+1}-1)v_k - t_{k+2}v_{k+1}\rangle - h^3 t_{k+1}t_{k+2}\langle v_{k+1}, e_k\rangle\\
&= h^3 t_{k+1}\Big(\langle \nabla f(y_k), -h(t_{k+1}-1)\nabla f_{k-1}(y_{k-1}) + h(t_{k+1}+t_{k+2}-1)\nabla f_k(y_k)\rangle\Big) - h^3 t_{k+1}t_{k+2}\langle v_{k+1}, e_k\rangle\\
&= h^4 t_{k+1}\Big(\langle \nabla f(y_k), -(t_{k+1}-1)\nabla f(y_{k-1}) + (t_{k+1}+t_{k+2}-1)\nabla f(y_k)\rangle\Big)\\
&- h^3 t_{k+1}t_{k+2}\langle v_{k+1}, e_k\rangle + h^4 t_{k+1}(t_{k+1}+t_{k+2}-1)\langle \nabla f(y_k), e_k\rangle - h^4 t_{k+1}(t_{k+1}-1)\langle \nabla f(y_k), e_{k-1}\rangle.
\end{aligned}$$

In addition

$$\frac{1}{2}h^4t_{k+1}^2\|\nabla f_k(y_k)\|^2-h^4t_{k+1}(t_{k+2}+t_{k+1}-1)\|\nabla f_k(y_k)\|^2=-\frac{1}{2}h^4t_{k+1}(2t_{k+2}+t_{k+1}-2)\|\nabla f_k(y_k)\|^2.$$

Collecting the last three estimates and applying the inequalities to (39), we obtain

$$\begin{aligned}
&E_{k+1}-E_k+h^2\Big(t_{k+1}^2-t_{k+2}^2+t_{k+2}\Big)(f(y_k)-\min f)\\
&\leq -\frac{h^2}{2L}(t_{k+1}-1)t_{k+1}\|\nabla f(y_k)-\nabla f(y_{k-1})\|^2\\
&+h^4t_{k+1}\Big(\langle\nabla f(y_k),-(t_{k+1}-1)\nabla f(y_{k-1})+(t_{k+1}+t_{k+2}-1)\nabla f(y_k)\rangle\Big)\\
&-\frac{1}{2}h^4t_{k+1}(2t_{k+2}+t_{k+1}-2)\|\nabla f(y_k)+e_k\|^2\\
&-h^2t_{k+1}\langle e_k,y_k-x^\star\rangle-h^3t_{k+1}t_{k+2}\langle v_{k+1},e_k\rangle\\
&+h^4t_{k+1}(t_{k+1}+t_{k+2}-1)\langle\nabla f(y_k),e_k\rangle-h^4t_{k+1}(t_{k+1}-1)\langle\nabla f(y_k),e_{k-1}\rangle.
\end{aligned}$$

After developing the expression $\|\nabla f(y_k)+e_k\|^2$, we arrive at

$$\begin{aligned}
&E_{k+1}-E_k+h^2\Big(t_{k+1}^2-t_{k+2}^2+t_{k+2}\Big)(f(y_k)-\min f)\\
&\leq -\frac{h^2}{2L}(t_{k+1}-1)t_{k+1}\|\nabla f(y_k)-\nabla f(y_{k-1})\|^2\\
&+h^4t_{k+1}\Big(\langle\nabla f(y_k),-(t_{k+1}-1)\nabla f(y_{k-1})+(t_{k+1}+t_{k+2}-1)\nabla f(y_k)\rangle\Big)\\
&-\frac{1}{2}h^4t_{k+1}(2t_{k+2}+t_{k+1}-2)\Big(\|\nabla f(y_k)\|^2+\|e_k\|^2+2\langle\nabla f(y_k),e_k\rangle\Big)\\
&-h^2t_{k+1}\langle e_k,y_k-x^\star\rangle-h^3t_{k+1}t_{k+2}\langle v_{k+1},e_k\rangle\\
&+h^4t_{k+1}(t_{k+1}+t_{k+2}-1)\langle\nabla f(y_k),e_k\rangle-h^4t_{k+1}(t_{k+1}-1)\langle\nabla f(y_k),e_{k-1}\rangle.
\end{aligned}$$

Taking the expectation conditionally on $\mathcal{F}_k$ and using conditional unbiasedness in (K_2^+), we get that ($\mathbb{P}$-a.s.)

$$\begin{aligned}
&\mathbb{E}\left[E_{k+1}\mid\mathcal{F}_k\right]-E_k+h^2\Big(t_{k+1}^2-t_{k+2}^2+t_{k+2}\Big)(f(y_k)-\min f)\\
&\leq -\frac{h^2}{2L}(t_{k+1}-1)t_{k+1}\|\nabla f(y_k)-\nabla f(y_{k-1})\|^2\\
&+h^4t_{k+1}\Big(\langle\nabla f(y_k),-(t_{k+1}-1)\nabla f(y_{k-1})+(t_{k+1}+t_{k+2}-1)\nabla f(y_k)\rangle\Big)\\
&-\frac{1}{2}h^4t_{k+1}(2t_{k+2}+t_{k+1}-2)\|\nabla f(y_k)\|^2-\frac{1}{2}h^4t_{k+1}(2t_{k+2}+t_{k+1}-2)\sigma_k^2\\
&+h^3t_{k+1}t_{k+2}\mathbb{E}\left[\|v_{k+1}\|^2\mid\mathcal{F}_k\right]^{1/2}\sigma_k,
\end{aligned}$$

where we used Cauchy-Schwartz inequality in the last term. Now we rely on Theorem 3.1, and in particular on (26) and (30) to infer that

$$\begin{aligned}
\|v_{k+1}\|=\frac{1}{h}\|y_{k+1}-y_k\|&\leq\frac{1}{h}\|y_{k+1}-x_{k+1}\|+\frac{1}{h}\|x_{k+1}-x_k\|+\frac{1}{h}\|x_k-y_k\|\\
&\leq\frac{2}{h}\|x_{k+1}-x_k\|+\frac{1}{h}\|x_k-x_{k-1}\|=o\left(\frac{1}{t_{k+1}}\right)+o\left(\frac{1}{t_k}\right)=o\left(\frac{1}{t_{k+1}}\right)\quad(\mathbb{P}\text{-a.s.}).
\end{aligned}$$

In the last equality we used again that (K_1^+) implies $t_{k+1} \leq 2t_k$. Therefore, there exists a nonnegative random variable η with $\operatorname{ess\,sup} \eta < +\infty$ such that $\mathbb{E}\left[\|v_{k+1}\|^2 \mid \mathcal{F}_k\right]^{1/2} \leq \eta/t_{k+1}$, and in turn

$$\begin{aligned}
&\mathbb{E}\left[E_{k+1} \mid \mathcal{F}_k\right] - E_k + h^2\Big(t_{k+1}^2 - t_{k+2}^2 + t_{k+2}\Big)(f(y_k) - \min f) \\
&\leq -\frac{h^2}{2L}(t_{k+1}-1)t_{k+1}\|\nabla f(y_k) - \nabla f(y_{k-1})\|^2 \\
&+ h^3 t_{k+1}\Big(\langle \nabla f(y_k), -h(t_{k+1}-1)\nabla f(y_{k-1}) + h(t_{k+1}+t_{k+2}-1)\nabla f(y_k)\rangle\Big) \\
&- \frac{1}{2}h^4 t_{k+1}(2t_{k+2}+t_{k+1}-2)\|\nabla f(y_k)\|^2 + 4\eta h^3 t_k \sigma_k,
\end{aligned}$$

where we used again that $t_{k+2} \leq 4t_k$ and we discarded the term involving σ_k^2 since $t_k \geq 1$ and thus $2t_{k+2} + t_{k+1} - 2 \geq 1$. Equivalently,

$$\mathbb{E}\left[E_{k+1} \mid \mathcal{F}_k\right] - E_k + h^2\Big(t_{k+1}^2 - t_{k+2}^2 + t_{k+2}\Big)(f(y_k) - \min f) \leq -R(\nabla f(y_{k-1}), \nabla f(y_k)) + 4\eta h^3 t_k \sigma_k, \tag{40}$$

where R is the quadratic form

$$\begin{aligned}
R(X,Y) = \frac{h^2}{2L}(t_{k+1}-1)t_{k+1}\|Y-X\|^2 + \frac{1}{2}h^4 t_{k+1}(2t_{k+2}+t_{k+1}-2)\|Y\|^2 \\
- h^3 t_{k+1}\left(\langle Y, -h(t_{k+1}-1)X + h(t_{k+1}+t_{k+2}-1)Y\rangle\right).
\end{aligned} \tag{41}$$

To conclude, we just need to prove that R is nonnegative. A standard procedure consists in computing a lower-bound $\min_X R(X,Y)$ for fixed Y. By taking the derivative of R with respect to X, we obtain that the minimum is achieved at $\bar{X}$ with $\bar{X} - Y = -h^2 L Y$. Therefore,

$$\begin{aligned}
\min_X R(X,Y) = \frac{h^2 L}{2}(t_{k+1}-1)t_{k+1}h^4\|Y\|^2 + \frac{1}{2}h^4 t_{k+1}(2t_{k+2}+t_{k+1}-2)\|Y\|^2 \\
- h^3 t_{k+1}\left(\langle Y, -h(t_{k+1}-1)(1-h^2L)Y + h(t_{k+1}+t_{k+2}-1)Y\rangle\right).
\end{aligned}$$

After reduction, we get

$$\min_X R(X,Y) = \frac{h^4 t_{k+1}}{2}\left((t_{k+1}-1)(2-h^2L) - 1\right)\|Y\|^2. \tag{42}$$

According to assumption (K_1^+), the coefficient of $f(y_k) - \min f$ in (40) is positive. We therefore discard this term in the rest of the proof. Combining (42) with (40), we obtain

$$\mathbb{E}\left[E_{k+1} \mid \mathcal{F}_k\right] - E_k \leq -\frac{h^4 t_{k+1}}{2}\left((t_{k+1}-1)(2-h^2L) - 1\right)\|\nabla f(y_k)\|^2 + 4\eta h^3 t_k \sigma_k.$$

Since $h^2 \in]0, 1/L]$ and $t_k \geq 1$, this can also be bounded as

$$\begin{aligned}
\mathbb{E}\left[E_{k+1} \mid \mathcal{F}_k\right] &\leq E_k - \frac{h^2 t_{k+1}}{2L}\left((t_{k+1}-1)(2-h^2L) - 1\right)\|\nabla f(y_k)\|^2 + \frac{4\eta h}{L}t_k\sigma_k \\
&\leq E_k - \frac{h^2 t_{k+1}(t_{k+1}-2)}{2L}\|\nabla f(y_k)\|^2 + \frac{4\eta h}{L}t_k\sigma_k \\
&= E_k - \frac{h^2 t_{k+1}^2}{2L}\|\nabla f(y_k)\|^2 + \frac{h^2 t_{k+1}}{L}\|\nabla f(y_k)\|^2 + \frac{4\eta h}{L}t_k\sigma_k \\
&\leq E_k - \frac{h^2 t_{k+1}^2}{2L}\|\nabla f(y_k)\|^2 + 2h^2 t_{k+1}(f(y_k) - \min f) + \frac{4\eta h}{L}t_k\sigma_k,
\end{aligned}$$

where we used co-coercivity of ∇f in the last inequality. The summability assumption in (K_2^+) together with the summability result in Theorem 3.2 allow then to invoke Lemma A.1 to get the claim. Observe that this also gives that E_k converges ($\mathbb{P}$-a.s.) to a non-negative valued random variable. □

Remark 3.2 Since $t_k \geq 1$, a direct consequence of the gradient summability shown in Theorem 3.3 is that the gradient sequence $(\nabla f(y_k))_{k\in\mathbb{N}}$ tends to zero ($\mathbb{P}$-a.s.) at least as quickly as at the rate $o(1/t_k)$. Observe also that this analysis gives another proof for the fast convergence of the function values (just carry on the proof starting from (40) without discarding the term involving the function values).

Note that the above proof has been notably simplified by using the conclusions already obtained in Theorem 3.2, and in particular to properly bound the terms involving v_{k+1} (which are not in $\mathcal{F}_k$). Extending this proof to the case where the step-size s_k is varying appears to be straightforward, but comes at the price of tedious and longer computations. We avoid this for the sake of brevity.

4 Discussion of Particular Parameter Choices

Let consider the theoretical guarantees obtained under the condition that there exists $c \in [0, 1[$ such that, for every $k \geq 1$

$$\frac{1}{1-\alpha_{k+1}} - \frac{1}{1-\alpha_k} \leq c. \tag{43}$$

This implies some important properties of t_k. One significant observation is a trade-off between stability to errors and fast convergence of $((\mathrm{SRAG})_{\alpha_{k+1}})$. Some choices of α_k will be less stringent on the required summability of the error variance for convergence, but will result in slower convergence rate and vice-versa.

In presenting the details, let us start with the following results that were obtained in [5, Proposition 3.3, 3.4]. The first one presents some general conditions on (α_k) and c that ensure the satisfaction of (K_0) and (K_1) (resp. (K_1^+)). The second one provides an explicit expression of t_k as a function of α_k.

Proposition 4.1 *Let $c \in [0, 1[$ and let $(\alpha_k)_{k\in\mathbb{N}}$ be a sequence satisfying $\alpha_k \in [0, 1[$ together with inequality (43) for every $k \geq 1$. Then condition (K_0) is satisfied. Moreover, we have for every $k \geq 1$,*

$$t_{k+1} \leq \frac{1}{(1-c)(1-\alpha_k)}.$$

If $c \leq 1/3$ (resp. $c < 1/3$), then condition (K_1) (resp. (K_1^+)) is fulfilled.

Proposition 4.2 *Let $(\alpha_k)_{k\in\mathbb{N}}$ be a sequence such that $\alpha_k \in [0, 1[$ for every $k \geq 1$. Given $c \in [0, 1[$, assume that*

$$\lim_{k\to+\infty} \frac{1}{1-\alpha_{k+1}} - \frac{1}{1-\alpha_k} = c.$$

Then, we have

$$t_{k+1} \sim \frac{1}{(1-c)(1-\alpha_k)} \quad \text{as } k \to +\infty.$$

Let us now consider several possible iterative regimes defining α_k.

4.1 Case 1: $\alpha_k = 1 - \dfrac{\alpha}{k}$, $\alpha > 0$:

This corresponds to the choice made in the (deterministic) Nesterov and Ravine methods studied in [10]. In this case, for every $k \geq 1$,

$$\frac{1}{1-\alpha_{k+1}} - \frac{1}{1-\alpha_k} = \frac{k+1}{\alpha} - \frac{k}{\alpha} = \frac{1}{\alpha}.$$

Therefore, condition (43) is satisfied with $c = \frac{1}{\alpha}$. If $\alpha \geq 3$ (resp. $\alpha > 3$), we have $c \in]0, 1/3]$ (resp. $c \in]0, 1/3[$). According to Proposition 4.2, we have for every $k \geq 1$,

$$t_{k+1} \sim \frac{1}{(1-c)(1-\alpha_k)} = \frac{\alpha}{\alpha - 1}\,\frac{k}{\alpha} = \frac{k}{\alpha - 1}.$$

Indeed, one can easily show that the equality $t_{k+1} = \frac{k}{\alpha - 1}$ is satisfied. Moreover,

$$t_{k+1}/t_k^2 = k(\alpha-1)/(k-1)^2 \geq (\alpha-1)/(k-1) \Rightarrow \sum_{k\in\mathbb{N}} \frac{t_{k+1}}{t_k^2} = +\infty.$$

Thus, specializing Theorem 3.2 and Theorem 3.3, we obtain the following statement.

Corollary 4.1 *Assume that* (H) *holds. Let* $(y_k)_{k\in\mathbb{N}}$ *be the sequence generated by* $(\mathrm{SRAG})_{\alpha_{k+1}}$ *with* $\alpha_k = 1 - \frac{\alpha}{k}$ *where* $\alpha > 3$*, and* $s_k \in]0, 1/L]$ *is a non-increasing sequence. Assume that*

$$\mathbb{E}\left[e_k \mid \mathcal{F}_k\right] = 0 \ \ (\mathbb{P}\text{-}a.s.) \quad \text{and} \quad (k s_k \sigma_k)_{k\in\mathbb{N}} \in \ell^1_+(\mathscr{F}).$$

Then, the following holds ($\mathbb{P}$*-a.s.*)*:*

(i) $f(y_k) - \min_{\mathcal{H}} f = o\left(\frac{1}{s_k k^2}\right)$ *and* $\|y_k - y_{k-1}\| = o\left(\frac{1}{k}\right)$ *;*

(ii) $\sum_{k\in\mathbb{N}} k s_k (f(y_k) - \min_{\mathcal{H}} f) < +\infty$ *and* $\sum_{k\in\mathbb{N}} k\|y_k - y_{k-1}\|^2 < +\infty$ *;*

(iii) If moreover $\inf_k s_k > 0$*, then* $\sum_{k\in\mathbb{N}} k^2 \|\nabla f(y_k)\|^2 < +\infty$ *and* $(y_k)_{k\in\mathbb{N}}$ *converges weakly* ($\mathbb{P}$*-a.s.*) *to an* $\operatorname{argmin}(f)$*-valued random variable.*

Another possible choice would be $\alpha_k = \frac{k}{k+\alpha}$ in which case we obtain exactly the same results as in Corollary 4.1. This corresponds to the popular choice of the the Nesterov extrapolation parameter. For $(\mathrm{SNAG})_{\alpha_k}$ with this choice of α_k, we recover and complete the results obtained in the literature; see e.g., [3,2,27,25].

4.2 Case 2: $\alpha_k = 1 - \dfrac{\alpha}{k^r}$, $\alpha > 0$, $r \in]0, 1[$:

In this case, we have

$$\frac{1}{1-\alpha_{k+1}} - \frac{1}{1-\alpha_k} = \frac{1}{\alpha}(k+1)^r - \frac{1}{\alpha}k^r = \frac{k^r}{\alpha}\left((1+1/k)^r - 1\right) \sim \frac{r}{\alpha}\,k^{r-1} \to 0 \quad \text{as } k \to +\infty.$$

For each $c > 0$, the condition $1/(1-\alpha_{k+1}) - 1/(1-\alpha_k) \leq c$ is satisfied for k large enough. On the other hand, we deduce from Proposition 4.2 that $t_k \sim \dfrac{k^r}{\alpha}$ as $k \to +\infty$. This implies that $\sum_{i=1}^{k} t_i \sim \dfrac{1}{\alpha(1+r)}\,k^{1+r}$ as $k \to +\infty$. Theorem 3.2 and Theorem 3.3 under this specification yields the following result.

Corollary 4.2 *Assume that* (H) *holds. Let* $(y_k)_{k\in\mathbb{N}}$ *be the sequence generated by* $(\mathrm{SRAG})_{\alpha_{k+1}}$ *with* $\alpha_k = 1 - \frac{\alpha}{k^r}$ *where* $\alpha > 0$ *and* $r \in]0,1[$*, and* $s_k \in]0, 1/L]$ *is an non-increasing sequence. Assume that*

$$\mathbb{E}\left[e_k \mid \mathcal{F}_k\right] = 0 \ (\mathbb{P}\text{-}a.s.) \quad \text{and} \quad \left(k^r s_k \sigma_k\right)_{k\in\mathbb{N}} \in \ell^1_+(\mathscr{F}).$$

Then, the following holds ($\mathbb{P}$-*a.s.*)*:*

(i) $f(y_k) - \min_{\mathcal{H}} f = o\left(\frac{1}{s_k k^{2r}}\right)$ *and* $\|y_k - y_{k-1}\| = o\left(\frac{1}{k^r}\right)$ *;*

(ii) $\sum_{k\in\mathbb{N}} k^r s_k (f(y_k) - \min_{\mathcal{H}} f) < +\infty$ *and* $\sum_{k\in\mathbb{N}} k^r \|y_k - y_{k-1}\|^2 < +\infty$ *;*

(iii) If moreover $\inf_k s_k > 0$*, then* $\sum_{k\in\mathbb{N}} k^{2r} \|\nabla f(y_k)\|^2 < +\infty$ *and* $(y_k)_{k\in\mathbb{N}}$ *converges weakly* ($\mathbb{P}$-*a.s.*) *to an* $\operatorname{argmin}(f)$*-valued random variable.*

It is clear from this result that this choice of α_k allows for a less stringent summability condition on the stochastic errors, but this comes at the price of a slower convergence rate. We are not aware of any such a result in the literature.

4.3 Case 3: α_k constant:

This corresponds to the choice made in the Polyak's heavy ball with friction method [33,35]. Since $\alpha_k \equiv \alpha \in [0,1[$ for every $k \geq 1$, condition (43) is clearly satisfied with $c = 0$. In turn, $t_k \equiv 1/(1-\alpha)$ for all $k \geq 1$. Applying Theorem 3.2 and Theorem 3.3 we get the following.

Corollary 4.3 *Assume that* (H) *holds. Let* $(y_k)_{k\in\mathbb{N}}$ *be the sequence generated by* $(\mathrm{SRAG})_{\alpha_{k+1}}$ *with* $\alpha_k \equiv \alpha \in [0,1[$*, and* $s_k \in]0, 1/L]$ *is an non-increasing sequence. Assume that*

$$\mathbb{E}\left[e_k \mid \mathcal{F}_k\right] = 0 \ (\mathbb{P}\text{-}a.s.) \quad \text{and} \quad \left(s_k \sigma_k\right)_{k\in\mathbb{N}} \in \ell^1_+(\mathscr{F}).$$

Then, the following holds ($\mathbb{P}$-*a.s.*)*:*

(i) $\sum_{k\in\mathbb{N}} s_k (f(y_k) - \min_{\mathcal{H}} f) < +\infty$ *and* $\sum_{k\in\mathbb{N}} \|y_k - y_{k-1}\|^2 < +\infty$ *;*

(ii) If moreover $\inf_k s_k > 0$*, then* $\sum_{k\in\mathbb{N}} \|\nabla f(y_k)\|^2 < +\infty$ *and* $(y_k)_{k\in\mathbb{N}}$ *converges weakly* ($\mathbb{P}$-*a.s.*) *to an* $\operatorname{argmin}(f)$*-valued random variable.*

For $(\mathrm{SNAG})_{\alpha_k}$ with this choice of α_k, we recover and complete the results obtained in the literature; see e.g., [41,18,29,14].

5 Conclusion

In this paper we studied the convergence properties of the stochastic Ravine optimization algorithm. We verified the intuition provided by recent analysis from the dynamics systems perspective showing that the Ravine and Nesterov accelerated gradient methods behave similarly, with identical convergence properties. Specifically, we showed that the same asymptotic guarantees as well as convergence rates apply with respect to function values, gradients and convergence of the iterates.

A Auxiliary lemmas

We here collect some important results that play a crucial role in the convergence analysis of $(\text{SNAG})_{\alpha_k}$.

Lemma A.1 (Convergence of non-negative almost supermartingales [36]) *Given a filtration $\mathscr{R} = (\mathcal{R}_k)_{k\in\mathbb{N}}$ and the sequences of real-valued random variables $(r_k)_{k\in\mathbb{N}} \in \ell_+(\mathscr{R})$, $(a_k)_{k\in\mathbb{N}} \in \ell_+(\mathscr{R})$, and $(z_k)_{k\in\mathbb{N}} \in \ell_+^1(\mathscr{R})$ satisfying, for each $k \in \mathbb{N}$*

$$\mathbb{E}\left[r_{k+1} \mid \mathcal{R}_k\right] - r_k \leq -a_k + z_k \ \ (\mathbb{P}\text{-}a.s.)$$

it holds that $(a_k)_{k\in\mathbb{N}} \in \ell_+^1(\mathscr{R})$ and $(r_k)_{k\in\mathbb{N}}$ converges ($\mathbb{P}$-a.s.) to a random variable valued in $[0, +\infty[$.

The following lemma is a consequence of Lemma A.1; see also the discussion in [36, Section 3].

Lemma A.2 *Given a filtration $\mathscr{R} = (\mathcal{R}_k)_{k\in\mathbb{N}}$, let the sequence of random variables $(\varepsilon_k)_{k\in\mathbb{N}} \in \ell_+(\mathscr{R})$ such that $\left(\left(\mathbb{E}\left[\varepsilon_k^2 \mid \mathcal{R}_{k-1}\right]\right)^{1/2}\right)_{k\in\mathbb{N}} \in \ell_+^1(\mathscr{R})$. Then*

$$\sum_{k\in\mathbb{N}} \varepsilon_k < +\infty \quad (\mathbb{P}\text{-}a.s.)\,.$$

Proof Let $\zeta_k = \varepsilon_k - \mathbb{E}\left[\varepsilon_k \mid \mathcal{R}_{k-1}\right]$ and $r_k = \left(\sum_{i=1}^k \zeta_i\right)^2$. We obviously have $\mathbb{E}\left[\zeta_{k+1} \mid \mathcal{R}_k\right] = 0$. Thus

$$\begin{aligned}\mathbb{E}\left[r_{k+1} \mid \mathcal{R}_k\right] &= \left(\sum_{i=1}^k \zeta_i\right)^2 + \sum_{i=1}^k \zeta_i \mathbb{E}\left[\zeta_{k+1} \mid \mathcal{R}_k\right] + \mathbb{E}\left[\zeta_{k+1}^2 \mid \mathcal{R}_k\right] \\ &= r_k + \mathbb{E}\left[\zeta_{k+1}^2 \mid \mathcal{R}_k\right] = r_k + \mathrm{Var}\left[\varepsilon_{k+1}^2 \mid \mathcal{R}_k\right] \leq r_k + \mathbb{E}\left[\varepsilon_{k+1}^2 \mid \mathcal{R}_k\right].\end{aligned}$$

It is easy to see that $\left(\left(\mathbb{E}\left[\varepsilon_k^2 \mid \mathcal{R}_{k-1}\right]\right)^{1/2}\right)_{k\in\mathbb{N}} \in \ell_+^1(\mathscr{R})$ implies $\left(\mathbb{E}\left[\varepsilon_k^2 \mid \mathcal{R}_{k-1}\right]\right)_{k\in\mathbb{N}} \in \ell_+^1(\mathscr{R})$, and we can apply Lemma (A.1) to get that

$$\lim_{k\to+\infty} r_k$$

exists and is finite ($\mathbb{P}$-a.s.). Using Jensen's inequality we have

$$0 \leq \sum_{i=1}^k \varepsilon_i = \sum_{i=1}^k \zeta_i + \sum_{i=1}^k \mathbb{E}\left[\varepsilon_i \mid \mathcal{R}_{i-1}\right] \leq r_k^{1/2} + \sum_{i=1}^k \left(\mathbb{E}\left[\varepsilon_i^2 \mid \mathcal{R}_{i-1}\right]\right)^{1/2}.$$

Passing to the limit using that $\left(\left(\mathbb{E}\left[\varepsilon_k^2 \mid \mathcal{R}_{k-1}\right]\right)^{1/2}\right)_{k\in\mathbb{N}} \in \ell_+^1(\mathscr{R})$ proves the claim. □

Lemma A.3 (Extended descent lemma) *Let $g : \mathcal{H} \to \mathbb{R}$ be a convex function whose gradient is L-Lipschitz continuous. Let $s \in]0, 1/L]$. Then for all $(x, y) \in \mathcal{H}^2$, we have*

$$g(y - s\nabla g(y)) \leq g(x) + \langle \nabla g(y),\, y - x\rangle - \frac{s}{2}\left\|\nabla g(y)\right\|^2 - \frac{s}{2}\left\|\nabla g(x) - \nabla g(y)\right\|^2. \tag{44}$$

See e.g. [7, Lemma 1]

B The Ravine method from a dynamic perspective

In this section, we consider the high resolution ODE of the Ravine method, and show that it exhibits damping governed by the Hessian. This will explain the fast convergence towards zero of the gradients satisfied by the Ravine method.

B.1 Dynamic tuning of the extrapolation coefficients

Let us first explain how to tune the extrapolation coefficients in the Ravine method, in order to obtain a dynamic interpretation of the algorithm. Critical to the understanding is the link between the Ravine method and the Nesterov method, as explained in Section 2, and the dynamic interpretation of the Nesterov method, due to Su, Boyd and Candès [39]. Consider the inertial gradient system

$$\ddot{x}(t) + \gamma(t)\dot{x}(t) + \nabla f(x(t)) = 0, \qquad ((\text{IGS})_\gamma)$$

which involves a general viscous damping coefficient $\gamma(\cdot)$. The implicit time discretization of $(\text{IGS})_\gamma$, with time step-size $h > 0$, $x_k = x(\tau_k)$, and $\tau_k = kh$ [3], gives

$$\frac{x_{k+1} - 2x_k + x_{k-1}}{h^2} + \gamma(kh)\frac{x_k - x_{k-1}}{h} + \nabla f(x_{k+1}) = 0.$$

Let $s := h^2$. After multiplication by s, we obtain

$$(x_{k+1} - x_k) - (x_k - x_{k-1}) + h\gamma(kh)(x_k - x_{k-1}) + s\nabla f(x_{k+1}) = 0. \tag{45}$$

Equivalently

$$x_{k+1} + s\nabla f(x_{k+1}) = x_k + (1 - h\gamma(kh))\,(x_k - x_{k-1}), \tag{46}$$

which gives

$$x_{k+1} = \operatorname{prox}_{sf}\left(x_k + (1 - h\gamma(kh))\,(x_k - x_{k-1})\right). \tag{47}$$

We obtain the inertial proximal algorithm

$$\begin{cases} y_k = x_k + (1 - h\gamma(kh))\,(x_k - x_{k-1}) \\ x_{k+1} = \operatorname{prox}_{sf}(y_k). \end{cases}$$

Following the general procedure described in [10], which consists in replacing the proximal step by a gradient step, we obtain $(\text{NAG})_{\alpha_k}$ with $\alpha_k = 1 - h\gamma(kh)$. Taking $\gamma(t) = \frac{\alpha}{t}$, we obtain $(\text{NAG})_{\alpha_k}$ with $\alpha_k = 1 - \frac{\alpha}{k}$, which provides fast convergence results. Observe that Algorithm $(\text{NAG})_{\alpha_k}$ makes sense for any arbitrarily given sequence of positive numbers $(\alpha_k)_{k\in\mathbb{N}}$. But for this algorithm to be directly connected by temporal discretization to the continuous dynamic $(\text{IGS})_\gamma$, it is necessary to take $\alpha_k = 1 - h\gamma(kh)$. Note that the case $\gamma(t) = \frac{\alpha}{t}$ is special, since due to the homogeneity property of $\gamma(\cdot)$, in this case α_k does not depend on h.

Let us now use the relations established in Section 2 between the Nesterov and the Ravine methods. Since $(x_k)_{k\in\mathbb{N}}$ satisfies $(\text{NAG})_{\alpha_k}$ with $\alpha_k = 1 - h\gamma(kh)$, we have that the associated sequence $(y_k)_{k\in\mathbb{N}}$ follows $(\text{RAG})_{\gamma_k}$ with $\gamma_k = \alpha_{k+1} = 1 - h\gamma((k+1)h)$.

B.2 High resolution ODE of the Ravine method

Let us now proceed with the high resolution ODE of the Ravine method $(\text{RAG})_{\gamma_k}$. The idea is not to let $h \to 0$, but to take into account the terms of order $h = \sqrt{s}$ in the asymptotic expansion, and to neglect the term of order $h^2 = s$. The high resolution method is extensively used in fluid mechanics, where physical phenomena occur at multiple scales. Indeed, by following an approach similar to that developed by Shi, Du, Jordan, and Su in [38], and Attouch and Fadili in [10], we are going to show that the Hessian-driven damping appears in the associated continuous inertial equation. Let us make this precise in the following result.

Theorem B.1 *The high resolution ODE with temporal step size $h = \sqrt{s}$ of the Ravine method $(\text{RAG})_{\gamma_k}$ with $\gamma_k = h\gamma((k+1)h)$ gives the inertial dynamic with Hessian driven damping*

$$\ddot{y}(t) + \gamma(t)\left(1 + \frac{\sqrt{s}}{2}\gamma(t)\right)\dot{y}(t) + \sqrt{s}\nabla^2 f(y(t))\dot{y}(t) + \left(1 + \frac{\sqrt{s}}{2}\gamma(t)\right)\nabla f(y(t)) = 0. \tag{48}$$

[3] We take the τ_k notation instead of the usual t_k, because t_k will be used with a different meaning, and it is used extensively in the paper.

Proof Set $\gamma_k = 1 - h\gamma((k+1)h)$. By definition of the Ravine method

$$y_{k+1} = y_k - s\nabla f(y_k) + \gamma_k\Big(y_k - s\nabla f(y_k)) - (y_{k-1} - s\nabla f(y_{k-1}))\Big).$$

Equivalently

$$(y_{k+1} - y_k) - (y_k - y_{k-1}) + (1-\gamma_k)(y_k - y_{k-1}) + s\nabla f(y_k) + s\gamma_k\Big(\nabla f(y_k) - \nabla f(y_{k-1})\Big) = 0.$$

After dividing by $s = h^2$, we obtain

$$\frac{y_{k+1} - 2y_k + y_{k-1}}{h^2} + (1-\gamma_k)\frac{y_k - y_{k-1}}{h^2} + \nabla f(y_k) + \gamma_k(\nabla f(y_k) - \nabla f(y_{k-1})) = 0. \tag{49}$$

Notice then that

$$\frac{y_k - y_{k-1}}{h^2} = \frac{y_{k+1} - y_k}{h^2} - \frac{y_{k+1} - 2y_k + y_{k-1}}{h^2}.$$

So, (49) can be formulated equivalently as follows

$$\gamma_k\frac{y_{k+1} - 2y_k + y_{k-1}}{h^2} + (1-\gamma_k)\frac{y_{k+1} - y_k}{h^2} + \nabla f(y_k) + \gamma_k(\nabla f(y_k) - \nabla f(y_{k-1})) = 0. \tag{50}$$

After dividing by γ_k, we get

$$\frac{y_{k+1} - 2y_k + y_{k-1}}{h^2} + \frac{1-\gamma_k}{h\gamma_k}\frac{y_{k+1} - y_k}{h} + \frac{1}{\gamma_k}\nabla f(y_k) + (\nabla f(y_k) - \nabla f(y_{k-1})) = 0. \tag{51}$$

Building on (51), we use Taylor expansions taken at a higher order (here, order four) than for the low resolution ODE. For each $k \in \mathbb{N}$, set $\tau_k = (k+c)h$, where c is a real parameter that will be adjusted further. Assume that $y_k = Y(\tau_k)$ for some smooth curve $\tau \mapsto Y(\tau)$ defined for $\tau \geq t_0 > 0$. Performing a Taylor expansion in powers of h, when h is close to zero, of the different quantities involved in (51), we obtain

$$y_{k+1} = Y(\tau_{k+1}) = Y(\tau_k) + h\dot{Y}(\tau_k) + \frac{1}{2}h^2\ddot{Y}(\tau_k) + \frac{1}{6}h^3\dddot{Y}(\tau_k) + \mathcal{O}(h^4) \tag{52}$$

$$y_{k-1} = Y(\tau_{k-1}) = Y(\tau_k) - h\dot{Y}(\tau_k) + \frac{1}{2}h^2\ddot{Y}(\tau_k) - \frac{1}{6}h^3\dddot{Y}(\tau_k) + \mathcal{O}(h^4). \tag{53}$$

By adding (52) and (53) we obtain

$$\frac{y_{k+1} - 2y_k + y_{k-1}}{h^2} = \ddot{Y}(\tau_k) + \mathcal{O}(h^2).$$

Moreover, (52) gives

$$\frac{y_{k+1} - y_k}{h} = \dot{Y}(\tau_k) + \frac{1}{2}h\ddot{Y}(\tau_k) + \mathcal{O}(h^2).$$

By Taylor expansion of ∇f we have

$$\nabla f(y_k) - \nabla f(y_{k-1}) = h\nabla^2 f(Y(\tau_k))\dot{Y}(\tau_k) + \mathcal{O}\left(h^2\right).$$

Plugging all of the above results into (51), we obtain

$$[\ddot{Y}(\tau_k) + \mathcal{O}(h^2)] + \frac{1-\gamma_k}{h\gamma_k}\left[\dot{Y}(\tau_k) + \frac{1}{2}h\ddot{Y}(\tau_k) + \mathcal{O}(h^2)\right] + \frac{1}{\gamma_k}\nabla f(Y(\tau_k)) + \left[h\nabla^2 f(Y(\tau_k))\dot{Y}(\tau_k) + \mathcal{O}\left(h^2\right)\right] = 0.$$

Multiplying by $\frac{h\gamma_k}{1-\gamma_k}$, we obtain in an equivalent way

$$\frac{h\gamma_k}{1-\gamma_k}\ddot{Y}(\tau_k) + \dot{Y}(\tau_k) + \frac{1}{2}h\ddot{Y}(\tau_k) + \frac{h}{1-\gamma_k}\nabla f(Y(\tau_k)) + \frac{h^2\gamma_k}{1-\gamma_k}\nabla^2 f(Y(\tau_k))\dot{Y}(\tau_k) + \mathcal{O}(h^3) = 0.$$

After reduction of the terms involving $\ddot{Y}(t_k)$, we obtain

$$\frac{h(1+\gamma_k)}{2(1-\gamma_k)}\ddot{Y}(\tau_k) + \dot{Y}(\tau_k) + \frac{h}{1-\gamma_k}\nabla f(Y(\tau_k)) + \frac{h^2\gamma_k}{1-\gamma_k}\nabla^2 f(Y(\tau_k))\dot{Y}(\tau_k) + \mathcal{O}(h^3) = 0.$$

Multiplication by $\frac{2(1-\gamma_k)}{h(1+\gamma_k)}$ then yields

$$\ddot{Y}(\tau_k) + \frac{2(1-\gamma_k)}{h(1+\gamma_k)}\dot{Y}(\tau_k) + \frac{2}{1+\gamma_k}\nabla f(Y(\tau_k)) + \frac{2h\gamma_k}{1+\gamma_k}\nabla^2 f(Y(\tau_k))\dot{Y}(\tau_k) + \mathcal{O}(h^2) = 0. \tag{54}$$

According to $\gamma_k = 1 - h\gamma((k+1)h)$, and $\tau_k = (k+1)h$, we obtain

$$\ddot{Y}(\tau_k) + \frac{\gamma(\tau_k)}{1-\frac{h}{2}\gamma(\tau_k)}\dot{Y}(\tau_k) + \frac{1}{1-\frac{h}{2}\gamma(\tau_k)}\nabla f(Y(\tau_k)) + h\frac{1-h\gamma(\tau_k)}{1-\frac{h}{2}\gamma(\tau_k)}\nabla^2 f(Y(\tau_k))\dot{Y}(\tau_k) + \mathcal{O}(h^2) = 0.$$

By neglecting the term of order $s = h^2$, and keeping the terms of order $h = \sqrt{s}$, we obtain the inertial dynamic with Hessian driven damping

$$\ddot{Y}(t) + \gamma(t)\left(1 + \frac{\sqrt{s}}{2}\gamma(t)\right)\dot{Y}(t) + \sqrt{s}\nabla^2 f(Y(t))\dot{Y}(t) + \left(1 + \frac{\sqrt{s}}{2}\gamma(t)\right)\nabla f(Y(t)) = 0.$$

This completes the proof. □

Remark B.1 The high resolution ODE of the Ravine method exhibits Hessian driven damping. In addition, it incorporates a gradient correcting term weighted with a coefficient of $\left(1 + \frac{\sqrt{s}}{2}\gamma(t)\right)$. This is in accordance with [10] and [38]. Surprisingly, there is also a correction which appears in the viscosity term, the coefficient $\left(1 + \frac{\sqrt{s}}{2}\gamma(t)\right)$ in front of the velocity. Indeed as we already observed, the Nesterov case is very specific. When $\gamma(t) = \frac{\alpha}{t}$, we have $s = 1 - h\gamma((k+1)h) = 1 - \frac{\alpha}{k+1}$. Returning to (54), we have

$$\frac{2(1-s)}{h(1+s)} = \frac{\alpha}{h(k+1-\frac{\alpha}{2})}.$$

Taking $\tau_k = h(k+1-\frac{\alpha}{2})$ gives $\gamma(\cdot)$ as the viscosity coefficient of the limit equation.

References

1. Adly, S., Attouch, H., Fadili, J.: Comparative analysis of accelerated gradient algorithms for convex optimization: High and super resolution ode approach. Optimization (2024)
2. Allen-Zhu, Z.: Katyusha: the first direct acceleration of stochastic gradient methods. J. Mach. Learn. Res. **18**, Paper No. 221, 51 (2017)
3. Assran, M., Rabbat, M.: On the convergence of nesterov's accelerated gradient method in stochastic settings. In: Proceedings of the 37th International Conference on Machine Learning, vol. 119, pp. 410–420 (2020)
4. Attouch, H., Cabot, A.: Asymptotic stabilization of inertial gradient dynamics with time-dependent viscosity. J. Differential Equations **263**(9), 5412–5458 (2017). DOI 10.1016/j.jde.2017.06.024. URL `https://doi.org/10.1016/j.jde.2017.06.024`
5. Attouch, H., Cabot, A.: Convergence rates of inertial forward-backward algorithms. SIAM J. Optim. **28**(1), 849–874 (2018). DOI 10.1137/17M1114739. URL `https://doi.org/10.1137/17M1114739`
6. Attouch, H., Cabot, A., Chbani, Z., Riahi, H.: Inertial forward-backward algorithms with perturbations: application to Tikhonov regularization. J. Optim. Theory Appl. **179**(1), 1–36 (2018). DOI 10.1007/s10957-018-1369-3. URL `https://doi.org/10.1007/s10957-018-1369-3`
7. Attouch, H., Chbani, Z., Fadili, J., Riahi, H.: First-order optimization algorithms via inertial systems with Hessian driven damping. Math. Program. **193**(1, Ser. A), 113–155 (2022). DOI 10.1007/s10107-020-01591-1. URL `https://doi.org/10.1007/s10107-020-01591-1`
8. Attouch, H., Chbani, Z., Peypouquet, J., Redont, P.: Fast convergence of inertial dynamics and algorithms with asymptotic vanishing viscosity. Math. Program. **168**(1-2, Ser. B), 123–175 (2018). DOI 10.1007/s10107-016-0992-8. URL `https://doi.org/10.1007/s10107-016-0992-8`
9. Attouch, H., Czarnecki, M.O.: Asymptotic control and stabilization of nonlinear oscillators with non-isolated equilibria. J. Differential Equations **179**(1), 278–310 (2002). DOI 10.1006/jdeq.2001.4034. URL `https://doi.org/10.1006/jdeq.2001.4034`
10. Attouch, H., Fadili, J.: From the ravine method to the Nesterov method and vice versa: a dynamical system perspective. SIAM J. Optim. **32**(3), 2074–2101 (2022). DOI 10.1137/22M1474357. URL `https://doi.org/10.1137/22M1474357`
11. Attouch, H., Fadili, J., Kungurtsev, V.: On the effect of perturbations in first-order optimization methods with inertia and Hessian driven damping. Evol. Equ. Control Theory **12**(1), 71–117 (2023). DOI 10.3934/eect.2022022. URL `https://doi.org/10.3934/eect.2022022`

12. Attouch, H., Peypouquet, J.: The rate of convergence of Nesterov's accelerated forward-backward method is actually faster than $1/k^2$. SIAM J. Optim. **26**(3), 1824–1834 (2016). DOI 10.1137/15M1046095. URL https://doi.org/10.1137/15M1046095
13. Aujol, J.F., Dossal, C.: Stability of over-relaxations for the forward-backward algorithm, application to FISTA. SIAM J. Optim. **25**(4), 2408–2433 (2015). DOI 10.1137/140994964. URL https://doi.org/10.1137/140994964
14. Defazio, A., Jelassi, S.: Adaptivity without compromise: a momentumized, adaptive, dual averaged gradient method for stochastic optimization. J. Mach. Learn. Res. **23**, Paper No. [144], 34 (2022). DOI 10.22405/2226-8383-2022-23-5-130-144. URL https://doi.org/10.22405/2226-8383-2022-23-5-130-144
15. Driggs, D., Ehrhardt, M.J., Schönlieb, C.B.: Accelerating variance-reduced stochastic gradient methods. Math. Program. **191**(2, Ser. A), 671–715 (2022). DOI 10.1007/s10107-020-01566-2. URL https://doi.org/10.1007/s10107-020-01566-2
16. Drori, Y., Teboulle, M.: Performance of first-order methods for smooth convex minimization: a novel approach. Math. Program. **145**(1-2, Ser. A), 451–482 (2014). DOI 10.1007/s10107-013-0653-0. URL https://doi.org/10.1007/s10107-013-0653-0
17. Frostig, R., R. Ge, S.K., Sidford, A.: Un-regularizing: approximate proximal point and faster stochastic algorithms for empirical risk minimization. In: Proceedings of the 32nd International Conference on Machine Learning, vol. 37, pp. 2540–2548 (2015)
18. Gadat, S., Panloup, F., Saadane, S.: Stochastic heavy ball. Electron. J. Stat. **12**(1), 461–529 (2018). DOI 10.1214/18-EJS1395. URL https://doi.org/10.1214/18-EJS1395
19. Gelfand, I., Tsetlin, M.: Printszip nelokalnogo poiska v sistemah avtomatich. Optimizatsii, Dokl. AN SSSR **137**, 295–298 (1961). (in Russian)
20. Ghadimi, S., Lan, G.: Accelerated gradient methods for nonconvex nonlinear and stochastic programming. Mathematical Programming **156**(1), 59–99 (2016). DOI 10.1007/s10107-015-0871-8. URL https://doi.org/10.1007/s10107-015-0871-8
21. Gupta, K., Siegel, J.W., Wojtowytsch, S.: Nesterov acceleration despite very noisy gradients. In: Advances in Neural Information Processing Systems (2024). URL https://arxiv.org/abs/2302.05515
22. Haraux, A., Jendoubi, M.A.: On a second order dissipative ODE in Hilbert space with an integrable source term. Acta Math. Sci. Ser. B (Engl. Ed.) **32**(1), 155–163 (2012). DOI 10.1016/S0252-9602(12)60009-5. URL https://doi.org/10.1016/S0252-9602(12)60009-5
23. Jain, P., Netrapalli, P., Kakade, S.M., Kidambi, R., Sidford, A.: Parallelizing stochastic gradient descent for least squares regression: mini-batching, averaging, and model misspecification. J. Mach. Learn. Res. **18**, Paper No. 223, 42 (2017)
24. Kim, D., Fessler, J.A.: Optimized first-order methods for smooth convex minimization. Math. Program. **159**(1-2, Ser. A), 81–107 (2016). DOI 10.1007/s10107-015-0949-3. URL https://doi.org/10.1007/s10107-015-0949-3
25. Laborde, M., Oberman, A.: A lyapunov analysis for accelerated gradient methods: from deterministic to stochastic case. In: S. Chiappa, R. Calandra (eds.) Proceedings of the Twenty Third International Conference on Artificial Intelligence and Statistics, Proceedings of Machine Learning Research, vol. 108, pp. 602–612. PMLR (2020). URL https://proceedings.mlr.press/v108/laborde20a.html
26. Lan, G.: An optimal method for stochastic composite optimization. Mathematical Programming **133**(1), 365–397 (2012). DOI 10.1007/s10107-010-0434-y. URL https://doi.org/10.1007/s10107-010-0434-y
27. Lan, G.: First-order and stochastic optimization methods for machine learning. Springer Series in the Data Sciences. Springer, Cham ([2020] ©2020). DOI 10.1007/978-3-030-39568-1. URL https://doi.org/10.1007/978-3-030-39568-1
28. Lin, H., Mairal, J., Harchaoui, Z.: Catalyst acceleration for first-order convex optimization: from theory to practice. J. Mach. Learn. Res. **18**, Paper No. 212, 54 (2017)
29. Loizou, N., Richtárik, P.: Momentum and stochastic momentum for stochastic gradient, Newton, proximal point and subspace descent methods. Comput. Optim. Appl. **77**(3), 653–710 (2020). DOI 10.1007/s10589-020-00220-z. URL https://doi.org/10.1007/s10589-020-00220-z
30. Nesterov, Y.: Introductory lectures on convex optimization, Applied Optimization, vol. 87. Kluwer Academic Publishers, Boston, MA (2004). DOI 10.1007/978-1-4419-8853-9. URL https://doi.org/10.1007/978-1-4419-8853-9. A basic course
31. Nesterov, Y.E.: A method for solving the convex programming problem with convergence rate $O(1/k^2)$. Dokl. Akad. Nauk SSSR **269**(3), 543–547 (1983)
32. Park, C., Park, J., Ryu, E.K.: Factor-$\sqrt{2}$ acceleration of accelerated gradient methods. Appl. Math. Optim. **88**(3), Paper No. 77, 38 (2023). DOI 10.1007/s00245-023-10047-9. URL https://doi.org/10.1007/s00245-023-10047-9
33. Poljak, B.T.: Some methods of speeding up the convergence of iterative methods. Ž. Vyčisl. Mat i Mat. Fiz. **4**, 791–803 (1964)
34. Polyak, B.: Accelerated gradient methods revisited. In: Workshop Variational Analysis and Applications. Erice (August 28-September 5, 2018)

35. Polyak, B.T.: Introduction to optimization. Translations Series in Mathematics and Engineering. Optimization Software, Inc., Publications Division, New York (1987). Translated from the Russian, With a foreword by Dimitri P. Bertsekas
36. Robbins, H.: Selected papers. Springer-Verlag, New York (1985). Edited and with a preface by T. L. Lai and D. Siegmund, With an interview of Herbert Robbins by Warren Page
37. Schmidt, M., Roux, N.L., Bach, F.: Convergence rates of inexact proximal-gradient methods for convex optimization. In: Proceedings of the 24th International Conference on Neural Information Processing Systems, pp. 1458–1466 (2011)
38. Shi, B., Du, S.S., Jordan, M.I., Su, W.J.: Understanding the acceleration phenomenon via high-resolution differential equations. Math. Program. **195**(1-2, Ser. A), 79–148 (2022). DOI 10.1007/s10107-021-01681-8. URL `https://doi.org/10.1007/s10107-021-01681-8`
39. Su, W., Boyd, S., Candès, E.J.: A differential equation for modeling Nesterov's accelerated gradient method: theory and insights. J. Mach. Learn. Res. **17**, Paper No. 153, 43 (2016)
40. Villa, S., Salzo, S., Baldassarre, L., Verri, A.: Accelerated and inexact forward-backward algorithms. SIAM J. Optim. **23**(3), 1607–1633 (2013). DOI 10.1137/110844805. URL `https://doi.org/10.1137/110844805`
41. Yan, B.: Theoretical Analysis for Convex and Non-Convex Clustering Algorithms. ProQuest LLC, Ann Arbor, MI (2018). URL `http://gateway.proquest.com/openurl?url_ver=Z39.88-2004&rft_val_fmt=info:ofi/fmt:kev:mtx:dissertation&res_dat=xri:pqm&rft_dat=xri:pqdiss:28166146`. Thesis (Ph.D.)–The University of Texas at Austin