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 $O(\cdot)$ to $o(\cdot)$ in 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 comprehen-

* 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

sive 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

Communicated by Radu Ioan Bot.

1 Introduction

Given a real Hilbert space $\mathcal{H}$, our work 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:

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

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 inertial/extrapolation/momentum parameters $(\gamma_k)_{k\in\mathbb{N}}$ 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, conditioned on y_k, e_k is a *zero-mean stochastic noise* term.

One of the motivations for this additive perturbation model comes from stochastic optimization, where the gradient may be evaluated only approximately, either because of physical, numerical or computational reasons. The prototype example we think of is the stochastic optimization problem of the form

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

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 application domains that include machine learning and signal processing. Computing $\nabla f(x)$ (or even $f(x)$) is either impossible, when F can only be sampled in an online streaming manner, or computationally very expensive, when it is the empirical risk across a very large number of data samples. 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 $e_k = \nabla f(x_k) - \widehat{\nabla f}(x_k)$. Denote $\mathcal{F}_k$ the sub-σ-algebra generated by $\sigma\left(x_i, e_{i-1} : i \leq k\right)$. Observe that conditioned on $\mathcal{F}_k$, and by drawing independently m_k samples of ξ at iteration k, one has

$$\begin{aligned} \mathbb{E}\left[e_k \mid \mathcal{F}_k\right] = 0 \quad \text{and} \quad \mathbb{E}\left[\left\|e_k\right\|^2 \mid \mathcal{F}_k\right] &= \frac{1}{m_k}\left(\mathbb{E}_\xi[\|\nabla F(x_k, \xi)\|^2 \mid \mathcal{F}_k] - \|\nabla f(x_k)\|^2\right) \\ &\leq \frac{1}{m_k}\mathbb{E}_\xi[\|\nabla F(x_k, \xi)\|^2 \mid \mathcal{F}_k]. \end{aligned}$$

Thus if the gradient is bounded almost surely, then

$$\mathbb{E}\left[\left\|e_k\right\|^2 \mid \mathcal{F}_k\right] = O(1/m_k).$$

The boundedness assumption on the gradient is quite common in the literature, and it holds for some popular functions such as the logistic loss, or if the sequence of iterates is assumed a priori to be almost surely bounded. Therefore to make this variance decay fast enough with k, which will be made precise in our analysis, one has to take m_k as increasing with k at a fast enough rate. We will show in particular that this decay rate will depend on the choice of the sequences s_k and γ_k and presents a trade-off between stability to stochastic perturbation and fast convergence of (RAG_{γ_k}); see the detailed description of our contributions in Section 1.2 and the thorough discussion after Theorem 3.1.

1.1 Previous Work

1.1.1 Exact Inertial Dynamics and Algorithms

Damped inertial dynamics have a natural mechanical and physical interpretation. Asymptotically, they tend to stabilize the system at a minimizer of the global energy function, that is mitigate oscillations of overshoot. As such, they offer an intuitive way to develop fast and efficient optimization methods. B. Polyak [36,37] initiated the use of inertial dynamics to accelerate the gradient method in optimization based on a second order in time inertial dynamical system with a fixed viscous damping coefficient; the so-called Heavy Ball with Friction (HBF) method. (HBF) provided momentum to the dynamics that accelerated the convergence for strongly convex objectives. A discretization of HBF, that we refer to as HBF algorithm, leads to the iterative scheme [36]

$$x_{k+1} = x_k + \beta_k\left(x_k - x_{k-1}\right) - s_k \nabla f(x_k). \tag{HBF_{β_k}}$$

With an appropriate and iteration-independent choice of s_k and $\beta_k > 0$ in the strongly convex case, that depends on the objective conditioning, (HBF_{β_k}) inherits the optimal linear convergence properties of the continuous dynamics.

The Ravine method (RAG$_{\gamma_k}$) was introduced by Gelfand and Tsetlin [17] 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 shown in Figure 1.

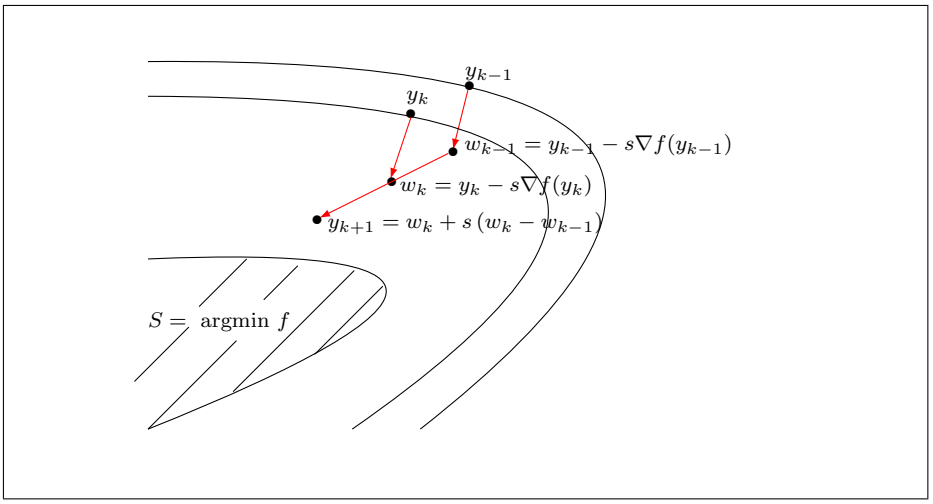

Fig. 1 Ravine Accelerated Gradient method.

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 [9,38,42]. It comes naturally into the picture when considering optimized first-order methods for smooth convex minimization, see [14,23,35].

When $\gamma_k = 1 - \frac{\alpha}{k}$, which, for $\alpha \geq 3$, the Ravine method (RAG$_{\gamma_k}$) is closely related with the Nesterov accelerated gradient (NAG for short) method (see (NAG$_{\alpha_k}$) [33,32]), 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 for each other in some of the literature. Their precise correspondence will be clearly elucidated in Section 2.

Notably, recent research concerning the understanding of accelerated first-order optimization methods, seen as temporal discretized dynamic systems, has made it possible to clarify the correspondence between these two methods; see the recent work by some of the authors in [9,1]. 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.

1.1.2 Inexact and Stochastic Inertial Dynamics

Due to the importance of the subject in optimization and control, several papers have been devoted to the study of perturbations in dissipative inertial systems and in the corresponding

accelerated first-order algorithms that can be modeled as these systems' dynamics' discretizations. This topic was first considered in the case of a fixed viscous damping and deterministic perturbations in [8,21]. Then it was studied as far as implications for the accelerated gradient method of Nesterov, from the standpoint of the corresponding inertial dynamics with vanishing viscous damping; see [42,5,7,12,40,45]. These results were extended to stochastic continuous time inertial dynamics in [31,30].

1.1.3 Stochastic Inertial Algorithms

Stochastic gradient descent methods with inertia/momentum are at the core of optimization subroutines in many data science applications such as machine learning and data processing. Rigorously speaking, these methods have provable acceleration over gradient descent only for convex objectives with exact gradients, or when the objective is in the form of a large finite sum of deterministic functions and the algorithm incorporates variance reduction techniques. In fact, with standard choices of step-size and inertia parameters, numerical experiments show that inertial algorithms may lose their superiority in convergence rate over simple gradient descent in the noisy setting, or can even diverge [15]. See also [22], who proved that for quadratic problems, there are situations in which it is impossible for any stochastic first-order oracle method to improve upon stochastic gradient descent (SGD). Yet, numerical studies have also shown that carefully tuned constant step-size and inertia momentum parameters can lead to improvements over stochastic gradient descent (SGD) in deep learning [44]. Overall, there has been a growing interest for obtaining convergence guarantees for stochastic extrapolation/momentum methods by considering narrowly particular parameter choices; see e.g., [25,18,27,20,22,2,46,13,34,16, 29,24,26,41,28]. This list is not exhaustive but is representative of most works related to ours.

Rates in expectation on the objective Many works provided guarantees of sublinear convergence rates in expectation on the objective for stochastic momentum/inertial methods. Most of these works assume (conditional) unbiasedness of the gradient and boundedness of its variance.

[25,18] developed an adaptation of (NAG_{α_k}) to stochastic (composite) convex optimization problems and obtained an optimal $O(1/\sqrt{k})$ rate for appropriate choice of the step-size and the extrapolation/momentum parameter; see also [26]. By relying on a Lyapunov function that dates back to the work of Attouch and co-authors, [24] obtain the convergence rate $O(1/\sqrt{k})$ for (SNAG_{α_k}) with non-vanishing noise and a step-size s_k decreasing sufficiently fast together with a specific choice of α_{k+1} tied to the step-size. An accelerated rate $O(1/k)$ was also obtained for the strongly convex case under appropriate choices for the parameters. [20] analyze (SNAG_{α_k}) with relative noise on the gradient, *i.e.* the noise variance is proportional to the squared magnitude of the actual gradient. For $\alpha_k = \frac{k}{k+3}$, *i.e.* the standard Nesterov scheme, they prove convergence at the same accelerated rate $O(1/k^2)$ as in the deterministic case, but only if the constant of proportionality in the noise is strictly less that 1.

[46] provided an analysis of the stochastic versions of (HBF_{β_k}) (that we coin SHBF) and (SNAG_{α_k}) with constant momentum/inertial parameters and asymptotically vanishing step-size $s_k = 1/\sqrt{k}$, under some boundedness assumptions of the gradient or its stochastic estimate. In the convex setting, they proved $O(1/\sqrt{k})$ ergodic (and not pointwise) convergence rates. See also [34] for similar results. [16] obtained a number of sublinear convergence guarantees for SHBF, including a $O(1/k^c)$ convergence rate, for any $c < 1$, in the strongly convex case when applying the decaying step-size $O(1/k^c)$ and an appropriately chosen inertial/memory parameter sequence.

In the strongly convex case, several works have studied convergence rates in expectation of stochastic inertial methods. [19] obtained an optimal $O(1/k)$ convergence rate under a boundedness assumption of the iterates. [2] showed a linear convergence rate of (SNAG_{α_k}) to a noise

dominated region[1] at the same accelerated rate as in the deterministic setting, provided that s_k and α_k are constant and appropriately chosen according to the objective condition number. [13] show that with a specific choice of constant parameters, the distribution of the iterates of (SNAG$_{\alpha_k}$) converges at an accelerated linear rate to a ball centered at a unique invariant distribution in the 1-Wasserstein metric, as long as the noise variance is small enough. They also prove linear convergence of the objective in expectation to a noise dominated region. For the specialized setting of a quadratic objective function, it has been shown that the SHBF iterates converge linearly at an accelerated rate, but only in expectation [29].

Almost sure guarantees While convergence guarantees in expectation are extensive, as reviewed above, the literature is much scarcer when it comes to almost sure guarantees, in particular for extrapolation/momentum methods when minimizing (1). This covers both convergence rates and more importantly, convergence of the sequence of iterates. The rare exceptions typically assume a finite sum objective and apply variance reduction, which as mentioned are beyond the scope of our work. For convergence theory that establishes guarantees in an almost sure sense, the Robbins-Siegmund lemma [39] (see Lemma A.1) turns out to be instrumental.

The almost sure convergence of SHBF to a local minimizer for sufficiently smooth coercive non-convex functions, verifying a Morse-Sard condition, was proven in [16] using the center stable manifold theorem, provided that the noise verifies an appropriate decrease and uniform ellipticity conditions so that SHBF is not trapped into spurious critical points.

By relying on an iterate averaging/memory viewpoint of SHBF, [41] showed that the function values converge almost surely at a rate close to $o(1/\sqrt{k})$ for an appropriate choice of parameters and under a slightly modified assumption of a bounded variance on the stochastic gradient evaluations. They have also claimed that the iterates of SHBF "converges to a minimizer"[2] almost surely. The derivation of almost sure convergence rates for SHBF heavily relied on the Lyapunov analysis of [11]. The work of [28] complemented the results of [41] by providing an almost sure convergence rate analysis for SHBF and (SNAG$_{\alpha_k}$) for strongly convex objectives, non-convex ones as well as general convex ones. They considered vanishing step-sizes and constant extrapolation/momentum parameters in $]0,1[$. For the convex case, as in our work, they provided an almost sure rate $O(1/k^{1/3-\epsilon})$ for s_k scaling as $1/k^{2/3+\epsilon}$, $\epsilon \in]0,1/3[$. They have also claimed almost sure convergence of the iterates of these algorithms.

1.2 Contributions

In this work, we propose and analyse both the stochastic Nesterov and Ravine methods (resp. (SNAG$_{\alpha_k}$) and (SRAG$_{\alpha_{k+1}}$)) with general extrapolation coefficient α_k for solving optimization problems of the form $(\mathcal{P})$ in an infinite-dimensional real separable Hilbert space 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, considering a general framework without narrow assumptions on α_k provides a broad picture of the convergence properties of this class of algorithms and reveals the precise role of α_k for balancing the trade-off between stability and fast convergence. Our contributions are the following:

[1] See Remark 3.2 and (7) for a definition.

[2] We take their terminology here, which is, strictly speaking, not rigorous. Indeed, as the iterates correspond to a random process, their limits are random variables and one should rather say that the sequence of iterates converges almost surely to a random variable taking values in the set of minimizers. From now on, we will use the latter.

- **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 range of values for the extrapolation sequence parameter.
- **Complexity estimates in expectation and almost sure sense**: we will establish fast convergence rates in expectation and in the almost sure sense, both in big-O and little-o, for the objective values, the gradient and the velocity.
- **Almost sure 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 precise impact of the extrapolation sequence on the convergence properties**: our results will highlight the trade-off between the choice of α_k, that of the step-sizes s_k, and the decrease of the error variance and their influence on the convergence of the values and gradients. In particular, some choices of the extrapolation parameter (and step-size sequence) 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 $\alpha_k = \frac{k}{k+\alpha}$ instead of $\alpha_k = 1 - \frac{\alpha}{k}$, as two of the many variations of the method.

1.3 Relation to Prior Work

Compared to all previous literature reviewed in Section 1.1.3 in the convex case, where each work essentially focuses on a specific algorithm with a certain choice of step-size and momentum parameters, our work is the first to deliver a unified analysis for both the stochastic Ravine and Nesterov algorithms with general parameters. Moreover, we are not aware of any work that tackles this problem in the infinite dimensional Hilbertian setting. Our proof relies a general Lyapunov analysis inspired by the one in [5] in the deterministic case. However, our extension from the deterministic to stochastic errors is quite involved and requires a novel careful analysis and several additional arguments, especially when it comes to establishing almost sure guarantees. Beyond a new Lyapunov analysis for properly isolating the moments of the martingale difference noise, we rely on the Robbins-Siegmund lemma (Lemma A.1) and its key consequence in Lemma A.2 for establishing the little-o guarantees of (SRAG$_{\alpha_{k+1}}$). To the best our knowledge, the latter is new and can be of independent interest. The proof of almost sure weak convergence of the iterates also necessitates a non-trivial density argument before applying Opial's lemma that relies on the separability of the Hilbert space.

Our convergence rates on the objective in expectation cover all those obtained in [25, 18, 27, 2, 46, 34, 16, 24, 26] as special cases for appropriate choice of the step-size and the inertial parameters. Our analysis does not require any boundedness assumption on the gradient or the noise, unlike for instance [46, 34].

Our convergence rate results are provided both in expectation and almost surely. While the former is standard in the literature, the latter is much less common, and the analysis less straightforward, as argued above. Notably, the almost sure weak convergence of the iterates is often overlooked by the overwhelming majority of existing works. In this respect, the only and closest work to ours is that of [41] and [28]. The algorithms analyzed in [41] and [28] are SHBF for the former, and SHBF and stochastic Nesterov's method with constant momentum for the latter. Here we analyze the stochastic Ravine and Nesterov's methods for general parameters.

Our setting allows for a broader class of flexible and physically meaningful (in terms of the corresponding ODE dynamics) choice of the parameters, as well as conditions on the noise variance, including constant variance. Although both works apply a Lyapunov analysis, ours is different and appears, in our view, to be more transparently informative. For the convergence rates, both in expectation and almost sure sense, we recover the same big-O and little-o rates on the objective as [41], and better than that of [28] (who did not provide little-o guarantees for the convex case), when we specify particular regimes for the parameters s_k and α_k. We have several additional and new results, including almost sure summability guarantees on the objective value, the velocity and the gradient. In our work (resp. [41,28]), almost sure (weak) convergence of the iterates of our algorithms (resp. theirs) to an $\operatorname{argmin}_{\mathcal{H}} f$-valued random variable is proved. However, we would like to point out that while the claim on almost sure convergence of the iterates in these two papers is true, a close inspection of their proofs shows that the arguments are unfortunately flawed. More precisely, these authors show that for any $x^\star \in \operatorname{argmin}_{\mathcal{H}} f$, $\lim_{k\to\infty} \|x_k - x^\star\|$ exists almost surely. However, the set of events of probability 1 on which this convergence event holds depends on $x^\star$. Thus, one cannot invoke Opial's Lemma directly, as they did, as the order of first order logical quantifiers present is inconsistent with their argument. To circumvent this difficulty, one has to use a non-trivial density argument (see the proof of Theorem 3.1(i)). Note that this is the only claim for which we need $\mathcal{H}$ to be separable in infinite dimension.

1.4 A Model Result

Taking $\alpha_{k+1} = 1 - \frac{\alpha}{k}$ in $(\mathrm{SRAG}_{\alpha_{k+1}})$ yields fast convergence the values and of the gradients towards zero at the same fast rate as in the deterministic case, provided that the noise decreases sufficiently fast. 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 $(w_i)_{i\le k-1}$. If $\alpha > 3$, $\mathbb{E}\left[e_k \mid \mathcal{F}_k\right] = 0$ 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 Corollary 4.1, 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, then the sequence $(y_k)_{k\in\mathbb{N}}$ converges weakly almost surely to a random variable valued in $\operatorname{argmin}_{\mathcal{H}} f$. Clearly, we recover the optimal convergence and summability rates known for the deterministic case [11] as soon as the noise variance vanishes fast enough. 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; see the discussion after each Corollary in Section 4 as well as Remark 3.1, Remark 3.2 and Remark 3.3.

1.5 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 From Nesterov to Ravine and Vice Versa

Let us first recall some basic facts concerning the NAG method. The latter with general extrapolation coefficients $(\alpha_k)_{k\in\mathbb{N}}$, as studied in [5], reads

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

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

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

and the fast convergence towards zero of the gradients (see [9])

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

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 [7,3,4,11,43].

The reason we used the subscript γ_k (resp. α_k) for the extrapolation coefficient in (RAG$_{\gamma_k}$) (resp. (NAG$_{\alpha_k}$)) was precisely to underline the difference and avoid confusion between the Ravine and Nesterov algorithms. A remarkable fact is that the variable y_k which enters the definition of (NAG$_{\alpha_k}$) follows the (RAG$_{\gamma_k}$) algorithm, with $\gamma_k = \alpha_{k+1}$. This generalizes the observation already made in [9] 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* (NAG$_{\alpha_k}$). *Then the associated sequence $(y_k)_{k\in\mathbb{N}}$ also follows the equations of the Ravine algorithm* (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 (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* (NAG$_{\alpha_k}$) *with $\alpha_k = \gamma_{k-1}$.*

Proof (i) Suppose that $(x_k)_{k\in\mathbb{N}}$ follows (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.*

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

(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

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

Equivalently

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

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 [9] 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 both the Ravine and Nesterov methods with stochastic errors in the evaluation of the gradients. We first start by proving the results for the Nesterov method before transferring them to the Ravine method thanks to Theorem 2.1. We will examine the fast convergence of the values and the convergence of iterates, and then we will 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 [10].

3.1 Values Convergence Rates and Convergence of the Iterates

We consider a stochastic version of $(\mathrm{SNAG}_{\alpha_k})$ 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 $k \in \mathbb{N}$,

$$\mathcal{F}_k := \sigma(x_i, e_{i-1} : i \leq k),$$

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 summarizes our main results.

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:

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

and

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

(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* $(\mathbb{P}\text{-}a.s.)$ *to an* $\mathrm{argmin}_{\mathcal{H}} f$*-valued random variable.*

The little-o statements of claim (ii) can be modified to get the same rate as in the deterministic case in [5, Theorem 3.4] but only at the price of a stronger summability assumption on the noise variance.

Before delving into the proof, a few remarks are in order. Overall, the discussion in these remarks will clearly illustrate the trade-off between the choice of α_k (hence t_k), that of the stepsizes s_k, and the decrease of the error variance and their role on the convergence rates of the values and gradients; see also Section 4.

Remark 3.1 (Vanishing noise variance) From claim (i), we have, for s_k constant and bounded away from 0, convergence at the rate $O(1/t_k^2)$ (resp. $O(\log(k)/t_k^2)$) in the objective in expectation if $(t_k\sigma_k)_{k\in\mathbb{N}} \in \ell^2_+(\mathscr{F})$ (resp. $t_k^2\mathbb{E}\left[\sigma_k^2\right] = O(1/k)$). For $t_k = (k-1)/(\alpha-1)$, $\alpha \geq 3$, which corresponds to the standard Ravine or Nesterov algorithms (see Section 4.1), one gets the accelerated rate $O(1/k^2)$ (in fact even $o(1/k^2)$ ($\mathbb{P}$-a.s.)), which is known to be optimal in the deterministic case. This is of course faster than the rate $O(1/k)$ one would get for SGD, but under the weaker assumption of $\left(\sqrt{k}\sigma_k\right)_{k\in\mathbb{N}} \in \ell^2_+(\mathscr{F})$.

If the above summability assumption is weakened to $(\sigma_k)_{k\in\mathbb{N}} \in \ell^2_+(\mathscr{F})$ (resp. $\mathbb{E}\left[\sigma_k^2\right] = O(1/k)$) such that σ_k is allowed to vanish at a slower rate, then the step-size must anneal at the rate $s_k \sim 1/t_k$ to compensate for this. The price to pay is then that the rate on the objective value in expectation goes from $O(1/t_k^2)$ to $O(1/t_k)$ (resp. from $O(\log(k)/t_k^2)$ to $O(\log(k)/t_k)$), and from $o(1/t_k^2)$ to $o(1/t_k)$ in the almost sure sense.

Remark 3.2 (Non-vanishing noise variance, vanishing step-size) Now consider non-vanishing noise with bounded variance (i.e. $\liminf_k \sigma_k > 0$ ($\mathbb{P}$-a.s.) and $\sup_k \mathbb{E}[\sigma_k^2] = B < \infty$). Consider the choice of $t_k \sim k^r$, where $r \in]1/2, 1]$ (see Section 4.1 and Section 4.2). Then setting the step-size to be $s_k = c/k^\delta$, with $c \in]0, 1/L]$ and $\delta \in]1, 2r[$, results in the convergence rates

$$\mathbb{E}\left[f(x_k) - \min f\right] = \begin{cases} O\left(1/k^{\delta-1}\right) & \delta \in]1, r+1/2[, \\ O\left(\log(k)/k^{r-1/2}\right) & \delta = r+1/2, \\ O\left(1/k^{2r-\delta}\right) & \delta \in]r+1/2, 2r[. \end{cases}$$

The choice $r = 1$ corresponds to the parameters in the standard Ravine or Nesterov algorithms (see Section 4.1), in which case one gets the convergence rates

$$\mathbb{E}\left[f(x_k) - \min f\right] = \begin{cases} O\left(1/k^{\delta-1}\right) & \delta \in]1, 3/2[, \\ O\left(\log(k)/\sqrt{k}\right) & \delta = 3/2, \\ O\left(1/k^{2-\delta}\right) & \delta \in]3/2, 2[. \end{cases}$$

The best rate approaches $O(1/\sqrt{k})$, obtained for $\delta = 3/2$. In fact, from our results, the rate even approaches $o(1/\sqrt{k})$ in the almost sure sense, similarly to the results in [41] and better than the near $O(1/k^{1/3})$ rate proved in [28].

On the other hand, for $r = 1$, if s_k only decreases as $1/t_k = O(1/k)$, then we have

$$\mathbb{E}\left[f(x_k) - \min f\right] = O\left(\frac{1}{k} + B\right). \tag{7}$$

This means that $\limsup_{k\to+\infty} \mathbb{E}\left[f(x_k) - \min f\right] \leq B$, *i.e.* $\mathbb{E}\left[f(x_k)\right]$ converges to a noisy region of size B around the minimal value (a.k.a. noise dominated region). However, one cannot say anything about the convergence of the iterates in this case.

Note that we cannot afford taking $r \in [0, 1/2]$, as otherwise, the bounds above both in big-O and little would be vacuous (in fact diverge). This means that for such parameter choice with non-vanishing noise variance, our convergence rates are not valid. However, one can still get convergence guarantees if the noise vanishes fast enough; see the discussion after Corollary 4.2 and Corollary 4.3.

Remark 3.3 (Ergodic convergence rate) Taking the full expectation in (19) and using Jensen's inequality, we have under (K_1^+), the following ergodic convergence rate in expectation

$$\mathbb{E}\left[f(\bar{x}_k) - \min f\right] \leq \frac{s_1 t_1^2 \mathbb{E}\left[f(x_0) - \min f\right] + \frac{1}{2}\mathbb{E}\left[\operatorname{dist}(x_0, S)^2\right] + 4\sum_{i=1}^k s_i^2 t_i^2 \mathbb{E}\left[\|e_i\|^2\right]}{(1-m)\sum_{i=1}^k s_i t_{i+1}},$$

where $\bar{x}_k = \sum_{i=1}^k s_i t_{i+1} x_i / (\sum_{i=1}^k s_i t_{i+1})$. For constant noise variance, taking $s_k = c/(\sqrt{k} t_{k+1})$, $c \in]0, 1/L]$, we get a rate $O(\log(k)/\sqrt{k})$. This generalizes the result in [46] beyond constant momentum parameter and without any boundedness assumption on the gradient. Taking $t_k \equiv 1$, *i.e.* $\alpha_k \equiv 0$, we recover the SGD algorithm for which get the known ergodic rate obtained with step-size sequence $c/\sqrt{k}$. Note that this ergodic rate was improved to $o(1/\sqrt{k})$ in the almost sure sense in [41].

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 $S = \operatorname{argmin}_{\mathcal{H}} f$. Define the sequence

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

Since f is convex and L-smooth, so is f_k. Let us apply (46) 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

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

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

Multiplying (8) by $t_{k+1} - 1$ (which is non-negative by definition), then adding the (9), we derive that

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

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 (10) and rearranging, we get

$$t_{k+1}(f_k(x_{k+1}) - f_k(x^\star)) \leq (t_{k+1}-1)(f_k(x_k) - f_k(x^\star)) + \langle \nabla f_k(y_k), z_k - x^\star\rangle - \frac{s_k}{2} t_{k+1}\left\|\nabla f_k(y_k)\right\|^2. \tag{11}$$

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

$$z_{k+1} - z_k = -s_k t_{k+1} \nabla f_k(y_k). \tag{12}$$

Thus

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

Dividing this by 2 and adding to (11), after multiplying the latter by $s_k t_{k+1}$, cancels all terms containing $\nabla f(y_k)$ and we arrive at

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

Let us take $x^\star$ as the closest point to z_k in S. Thus (13) is equivalent to

$$s_k t_{k+1}^2 (f_k(x_{k+1}) - f_k(x^\star)) + \frac{1}{2}\mathrm{dist}(z_{k+1}, S)^2 \leq s_k t_{k+1}(t_{k+1}-1)(f_k(x_k) - f_k(x^\star)) + \frac{1}{2}\mathrm{dist}(z_k, S)^2. \tag{14}$$

Let us now isolate the error terms. Inequality (14) is then equivalent to

$$\begin{aligned} s_k t_{k+1}^2 (f(x_{k+1}) - \min f) + \frac{1}{2}\mathrm{dist}(z_{k+1}, S)^2 \leq s_k t_{k+1}(t_{k+1}-1)(f(x_k) - \min f) + \frac{1}{2}\mathrm{dist}(z_k, S)^2 \\ - s_k \left\langle e_k,\, t_{k+1}^2 (x_{k+1} - x^\star) - t_{k+1}(t_{k+1}-1)(x_k - x^\star) \right\rangle . \end{aligned} \tag{15}$$

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, (15) 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

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

Taking the expectation conditionally on $\mathcal{F}_k$ in (16), we obtain

$$\mathbb{E}\left[V_{k+1} \mid \mathcal{F}_k\right] \leq V_k + s_k(t_{k+1}^2 - t_{k+1} - t_k^2)(f(x_k) - \min f) - s_k t_{k+1} \mathbb{E}\left[\left\langle e_k,\, z_{k+1} - x^\star \right\rangle \mid \mathcal{F}_k\right]. \tag{17}$$

We have

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

where we used (12) 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 (17) yields

$$\begin{aligned} \mathbb{E}\left[V_{k+1} \mid \mathcal{F}_k\right] &\leq V_k + s_k(t_{k+1}^2 - t_{k+1} - t_k^2)(f(x_k) - \min f) + s_k^2 t_{k+1}^2 \sigma_k^2 \\ &\leq V_k + s_k(t_{k+1}^2 - t_{k+1} - t_k^2)(f(x_k) - \min f) + 4 s_k^2 t_k^2 \sigma_k^2, \end{aligned} \tag{18}$$

where we used that assumption (K_1) implies $t_{k+1} \leq 2t_k$; see [5, Remark 3.3]. Using again (K_1), the second term in the rhs of (18) 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 (18) 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 (18) 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 (18), we get

$$\mathbb{E}\left[V_{k+1} \mid \mathcal{F}_k\right] \leq V_k - s_k(1-m)t_{k+1}(f(x_k) - \min f) + 4s_k^2 t_k^2 \sigma_k^2. \tag{19}$$

We can again invoke Lemma A.1 to get that

$$\sum_{k\geq 1} s_k t_{k+1}(f(x_k) - \min f) < +\infty \ (\mathbb{P}\text{-a.s.}). \tag{20}$$

Let

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

Combining [5, 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

$$\mathbb{E}\left[W_{k+1} \mid \mathcal{F}_k\right] \leq W_k - \frac{1-\alpha_k^2}{2}\|x_k - x_{k-1}\|^2 - s_k\mathbb{E}\left[\langle e_k,\, x_{k+1} - x_k\rangle \mid \mathcal{F}_k\right]. \tag{21}$$

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 (21) yields

$$\mathbb{E}\left[W_{k+1} \mid \mathcal{F}_k\right] \leq W_k - \frac{1-\alpha_k^2}{2}\|x_k - x_{k-1}\|^2 + s_k^2\sigma_k^2. \tag{22}$$

Multiplying (22) by t_{k+1}^2 and rearranging entails

$$\begin{aligned}
&\mathbb{E}\left[t_{k+1}^2 W_{k+1} \mid \mathcal{F}_k\right] \leq t_{k+1}^2 W_k - t_{k+1}^2\frac{1-\alpha_k^2}{2}\|x_k - x_{k-1}\|^2 + s_k^2 t_{k+1}^2\sigma_k^2 \\
&= t_k^2 W_k + s_k(t_{k+1}^2 - t_k^2)(f(x_k) - \min f) + \frac{t_{k+1}^2 - t_k^2 - t_{k+1}^2(1-\alpha_k^2)}{2}\|x_k - x_{k-1}\|^2 + s_k^2 t_{k+1}^2\sigma_k^2 \\
&\leq t_k^2 W_k + m s_k t_{k+1}(f(x_k) - \min f) - \frac{t_k}{2}\|x_k - x_{k-1}\|^2 + 4s_k^2 t_k^2\sigma_k^2.
\end{aligned} \tag{23}$$

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

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

since $t_k \geq 1$. We have already proved above (see (20)) 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 (23) to deduce that

$$\sum_{k\geq 1} t_k \left\|x_k - x_{k-1}\right\|^2 < +\infty \ (\mathbb{P}\text{-a.s.}). \tag{24}$$

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 (20) and (24) imply that

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

In turn

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

which entails 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 (23) 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 $S = \operatorname{argmin}_{\mathcal{H}} f$. 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 (ℙ-a.s.). For this, we use a standard argument that can be found e.g. in [11,5]. By [5, 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 $\sup_{k\in\mathbb{N}}\|x_k - x^\star\| \leq \eta < +\infty$ (ℙ-a.s.). Thus

$$h_{k+1} - h_k - \alpha_k(h_k - h_{k-1}) \leq \|x_k - x_{k-1}\|^2 + \eta s_k\|e_k\|\,. \tag{25}$$

Multiplying (25) 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 (ℙ-a.s.), we get using Lemma A.1 that $((h_k - h_{k-1})_+)_{k\in\mathbb{N}} \in \ell^1_+(\mathscr{F})$ (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 1 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

$$\lim_{k\to+\infty}\|x_k(\omega) - u_j\| = \tau_j(\omega), \quad \forall\omega \in \Omega_{u_j}. \tag{26}$$

Now, let $\omega \in \tilde{\Omega}$. Since $\tilde{\Omega} \subset \Omega_{u_j}$ for any $j \geq 1$, and using the triangle inequality and (26), 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 (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 $k\geq 1$

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

(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 (SRAG$_{\alpha_{k+1}}$). For this, we denote the canonical filtration associated to (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* (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* $\operatorname{argmin}_{\mathcal{H}} f$-*valued random variable.*

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

$$x_{k+1} = y_k - s_k(\nabla f(y_k)+e_k) \tag{27}$$

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

$$f(x_k)-\min f = o\left(\frac{1}{s_k t_k^2}\right) \quad \text{and} \quad \|x_k-x_{k-1}\| = o\left(\frac{1}{t_k}\right) \quad (\mathbb{P}\text{-a.s.}). \tag{28}$$

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

$$\sum_{k=1}^{+\infty} s_k t_k \|e_k\| < +\infty \quad (\mathbb{P}\text{-a.s.}), \tag{29}$$

and thus

$$s_k\|e_k\| = o\left(\frac{1}{t_k}\right) \quad (\mathbb{P}\text{-a.s.}). \tag{30}$$

Rearrange the terms in (27) 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

$$\begin{aligned} f(y_k) - \min_{\mathcal{H}} f &\leq f(x_k) - \min_{\mathcal{H}} f + \frac{1}{s_k} \langle x_{k+1} - y_k + s_k e_k, x_k - y_k \rangle \\ &\leq f(x_k) - \min_{\mathcal{H}} f + \frac{1}{s_k} \left(\|x_{k+1} - y_k\| + s_k \|e_k\| \right) \|x_k - y_k\| \\ &\leq f(x_k) - \min_{\mathcal{H}} f + \frac{1}{s_k} \left(\|x_{k+1} - x_k\| + \|x_k - y_k\| + s_k \|e_k\| \right) \|x_k - y_k\|. \quad (31) \end{aligned}$$

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

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

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

$$\|y_k - x_k\| \leq \|x_k - x_{k-1}\|. \quad (32)$$

Combining (28), (30), (31) and (32) 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.), (29) 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 (28) and (32), 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 $\operatorname{argmin}_{\mathcal{H}} f$-valued random variable. □

3.2 Fast Convergence of the Gradients

In this section, the previous results on the stochastic Ravine method (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[3].

[3] Observe that embarking from (8)-(9) and using the refined estimate in (46) 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 $k \in \mathbb{N}$

$$v_k := \frac{1}{h}(y_k - y_{k-1}), \tag{33}$$

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 $k \in \mathbb{N}$*

$$t_{k+1}(v_k + h\nabla f_{k-1}(y_{k-1})) - (t_k - 1)(v_{k-1} + h\nabla f_{k-2}(y_{k-2})) = -h(t_k - 1)\nabla f_{k-1}(y_{k-1}). \tag{34}$$

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

$$(v_k + h\nabla f_{k-1}(y_{k-1})) - \alpha_k(v_{k-1} + h\nabla f_{k-2}(y_{k-2})) = -h\alpha_k\nabla f_{k-1}(y_{k-1}). \tag{35}$$

In view of (6), the last equality is also equivalent to (34). 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

$$E_{k+1} - E_k \leq h^2(t_{k+1}-1)t_{k+1}(f(y_k) - f(y_{k-1}))$$
$$+ h^2\Big((t_{k+2}-1)t_{k+2} - (t_{k+1}-1)t_{k+1}\Big)(f(y_k) - \min f) + \frac{1}{2}\|z_{k+1} - x^\star\|^2 - \frac{1}{2}\|z_k - x^\star\|^2. \quad (36)$$

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

$$\frac{1}{2}\|z_{k+1} - x^\star\|^2 - \frac{1}{2}\|z_k - x^\star\|^2 = \langle z_{k+1} - z_k, z_{k+1} - x^\star\rangle - \frac{1}{2}\|z_{k+1} - z_k\|^2. \quad (37)$$

Recall the constitutive equation given by (34) that we write as follows

$$t_{k+2}(v_{k+1} + h\nabla f_k(y_k)) - (t_{k+1}-1)(v_k + h\nabla f_{k-1}(y_{k-1})) = -h(t_{k+1}-1)\nabla f_k(y_k). \quad (38)$$

Using successively the definition of z_k and (38), 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 (37), 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 (34) again, written as,

$$(t_{k+1}-1)(v_k + h\nabla f_{k-1}(y_{k-1})) = t_{k+2}(v_{k+1} + h\nabla f_k(y_k)) + h(t_{k+1}-1)\nabla f_k(y_k). \quad (39)$$

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 (36) we get

$$\begin{aligned}
&E_{k+1} - E_k \le h^2(t_{k+1}-1)t_{k+1}(f(y_k) - f(y_{k-1}))\\
&+ h^2\Big((t_{k+2}-1)t_{k+2} - (t_{k+1}-1)t_{k+1}\Big)(f(y_k) - \min f) + \frac{1}{2}h^4 t_{k+1}^2\|\nabla f_k(y_k)\|^2\\
&- h^2 t_{k+1}\left\langle \nabla f_k(y_k), y_k - x^\star + ht_{k+2}v_{k+1} + h^2(t_{k+2}+t_{k+1}-1)\nabla f_k(y_k)\right\rangle. \qquad (40)
\end{aligned}$$

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 (40), and using $\nabla f_k(y_k) = \nabla f(y_k) + e_k$, we get

$$\begin{aligned}
&E_{k+1} - E_k \le -h^2(t_{k+1}-1)t_{k+1}\Big(\langle \nabla f(y_k), y_{k-1} - y_k\rangle + \frac{1}{2L}\|\nabla f(y_k) - \nabla f(y_{k-1})\|^2\Big)\\
&+ h^2\Big((t_{k+2}-1)t_{k+2} - (t_{k+1}-1)t_{k+1}\Big)(f(y_k) - \min f) - h^2 t_{k+1}(f(y_k) - \min f)\\
&+ \frac{1}{2}h^4 t_{k+1}^2\|\nabla f_k(y_k)\|^2 - h^2 t_{k+1}\left\langle \nabla f_k(y_k), ht_{k+2}v_{k+1} + h^2(t_{k+2}+t_{k+1}-1)\nabla f_k(y_k)\right\rangle\\
&- h^2 t_{k+1}\langle y_k - x^\star, e_k\rangle. \qquad (41)
\end{aligned}$$

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 (39) 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 (41), 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 (28) and (32) to infer that ($\mathbb{P}$-a.s.),

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

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,

$$\mathbb{E}\left[E_{k+1} \mid \mathcal{F}_k\right] - E_k + h^2\Big(t_{k+1}^2 - t_{k+2}^2 + t_{k+2}\Big)(f(y_k) - \min f) \leq -R(\nabla f(y_{k-1}), \nabla f(y_k)) + 4\eta h^3 t_k \sigma_k, \tag{42}$$

where R is the quadratic form

$$\begin{aligned}
R(X,Y) = \frac{h^2}{2L}(t_{k+1}-1)t_{k+1}\|Y-X\|^2 + \frac{1}{2}h^4 t_{k+1}(2t_{k+2}+t_{k+1}-2)\|Y\|^2 \\
- h^3 t_{k+1}\left(\langle Y, -h(t_{k+1}-1)X + h(t_{k+1}+t_{k+2}-1)Y\rangle\right).
\end{aligned} \tag{43}$$

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

$$\min_X R(X,Y) = \frac{h^4 t_{k+1}}{2}\left((t_{k+1}-1)(2-h^2L) - 1\right)\|Y\|^2. \tag{44}$$

According to assumption (K_1^+), the coefficient of $f(y_k) - \min f$ in (42) is positive. We therefore discard this term in the rest of the proof. Combining (44) with (42), 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 3.4 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 (42) 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 $k \geq 1$

$$\frac{1}{1-\alpha_{k+1}} - \frac{1}{1-\alpha_k} \leq c. \tag{45}$$

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 [4, 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 (45) 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: $\alpha_k = 1 - \dfrac{\alpha}{k}$, $\alpha > 1$

This corresponds to the choice made in the (deterministic) Nesterov and Ravine methods studied in [9]. 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 (45) 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* $\operatorname{argmin}_{\mathcal{H}} f$*-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 Nesterov extrapolation parameter. For $(\mathrm{SNAG}_{\alpha_k})$ with this choice of α_k, we recover and complete some results obtained in the literature as discussed in Section 1.3, Remark 3.1 and Remark 3.2.

Recalling the example of risk minimization as (1) and the discussion thereafter, the summability assumption of Corollary 4.1 reads $\sum_{k\in\mathbb{N}} \frac{s_k k}{\sqrt{m_k}} < +\infty$, where m_k is the number of samples in the stochastic gradient estimate (2). Thus, choosing $m_k = O(k^{4+\delta})$, with $\delta > 0$, one can apply a constant step-size $s_k \equiv s \in]0, 1/L]$, and get, in the almost sure sense, guarantees of the fast convergence rate $o(1/k^2)$ on the objective and squared gradient as well as weak convergence of y_k to a random variable in the set of minimizers. The rate of the objective decrease becomes $o(1/k)$ if $m_k = O(k^{2+\delta})$ and $s_k = 1/(Lk)$, and $o(1/\sqrt{k})$ if $m_k = O(k^{1+\delta})$ and $s_k = 1/(Lk^{3/2})$.

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* (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 a non-increasing sequence. Assume that*

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

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* $\operatorname{argmin}_{\mathcal{H}} f$*-valued random variable.*

We are not aware of any result considering this case in the literature. See also the discussion in Remark 3.2 for the case of non-vanishing noise variance. On the other hand, the above result shows that this choice of α_k allows for a less stringent summability condition on the stochastic errors than the case of Section 4.1, but this comes at the price of a slower convergence rate. For instance, for the risk minimization example in (1), the summability assumption of Corollary 4.2 reads $\sum_{k\in\mathbb{N}} \frac{s_k k^r}{\sqrt{m_k}} < +\infty$. This entails the almost sure convergence rate $o(1/k^{2r})$ on the objective and squared gradient with constant step-size provided that $m_k = O(k^{4r+\delta})$ with δ such that $r + \delta/2 > 1$. The algorithm also exhibits almost sure weak convergence of y_k to a random variable in the set of minimizers.

4.3 Case 3: α_k Constant

This corresponds to the choice made in the Polyak's HBF method [36,37], though the algorithms are different (the gradient is not evaluated at the same point). Since $\alpha_k \equiv \alpha \in [0,1[$ for every $k \geq 1$, condition (45) 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* (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* $\operatorname{argmin}_{\mathcal{H}} f$-*valued random variable.*

Recall from Remark 3.2 that for constant noise variance and constant α_k, the (pointwise) big-O and little-o rates are vacuous. The reason is that the step-size will have to decrease fast enough to be summable, but this will hamper the convergence rates that scale as $1/s_k$. Note that one can nevertheless get an ergodic convergence rate $O(\log(k)/\sqrt{k})$ on the objective in expectation as discussed in Remark 3.3, hence recovering the result in [46].

For the risk minimization example in (1)–(2), the summability assumption of Corollary 4.3 reads as $\sum_{k\in\mathbb{N}} \frac{s_k}{\sqrt{m_k}} < +\infty$. Consider the case of vanishing noise variance, *i.e.* the sample size m_k increases with k. If $s_k \equiv s \in]0, 1/L]$ and $m_k = O(k^{2\delta})$, $\delta > 1$, then almost surely, one has $f(y_k) - \min_{\mathcal{H}} f \to 0$ and y_k converges weakly to a random variable in the set of minimizers. If m_k is allowed to increase at a slower rate, say $m_k = O(k^{1+\epsilon})$, $\epsilon > 0$, then s_k must decrease as $1/(L\sqrt{k})$ to ensure the error summability. But this would entail only that $\liminf_{k\to\infty} f(y_k) - \min_{\mathcal{H}} f \to 0$ almost surely. It is not clear what one could say about the convergence of the gradient and the iterates themselves in this situation.

A Auxiliary Lemmas

We here collect some important results that play a crucial role in the convergence analysis of (SNAG_{α_k}).

Lemma A.1 (Convergence of non-negative almost supermartingales [39]) *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* $k \in \mathbb{N}$

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

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

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

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

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

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*

$$g(y - s\nabla g(y)) \leq g(x) + \langle \nabla g(y),\, y - x\rangle - \frac{s}{2}\left\|\nabla g(y)\right\|^2 - \frac{s}{2}\left\|\nabla g(x) - \nabla g(y)\right\|^2. \tag{46}$$

See e.g. [6, 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 [43]. Consider the inertial gradient system

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

which involves a general viscous damping coefficient $\gamma(\cdot)$. The implicit time discretization of $(\mathrm{IGS})_\gamma$, with time step-size $h > 0$, $x_k = x(\tau_k)$, and $\tau_k = kh$ [4], 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

$$(x_{k+1} - x_k) - (x_k - x_{k-1}) + h\gamma(kh)(x_k - x_{k-1}) + s\nabla f(x_{k+1}) = 0. \tag{47}$$

Equivalently

$$x_{k+1} + s\nabla f(x_{k+1}) = x_k + (1 - h\gamma(kh))\,(x_k - x_{k-1}), \tag{48}$$

which gives

$$x_{k+1} = \mathrm{prox}_{sf}\left(x_k + (1 - h\gamma(kh))\,(x_k - x_{k-1})\right). \tag{49}$$

We obtain the inertial proximal algorithm

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

Following the general procedure described in [9], which consists in replacing the proximal step by a gradient step, we obtain $(\mathrm{NAG}_{\alpha_k})$ with $\alpha_k = 1 - h\gamma(kh)$. Taking $\gamma(t) = \frac{\alpha}{t}$, we obtain $(\mathrm{NAG}_{\alpha_k})$ with $\alpha_k = 1 - \frac{\alpha}{k}$, which provides fast convergence results. Observe that Algorithm $(\mathrm{NAG}_{\alpha_k})$ 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 $(\mathrm{IGS})_\gamma$, 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 $(\mathrm{NAG}_{\alpha_k})$ with $\alpha_k = 1 - h\gamma(kh)$, we have that the associated sequence $(y_k)_{k\in\mathbb{N}}$ follows $(\mathrm{RAG}_{\gamma_k})$ with $\gamma_k = \alpha_{k+1} = 1 - h\gamma((k+1)h)$.

[4] 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.

B.2 High Resolution ODE of the Ravine Method

Let us now proceed with the high resolution ODE of the Ravine method (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 [42], and Attouch and Fadili in [9], 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* (RAG$_{\gamma_k}$) *with $\gamma_k = h\gamma((k+1)h)$ gives the inertial dynamic with Hessian driven damping*

$$\ddot{y}(t) + \gamma(t)\left(1 + \frac{\sqrt{s}}{2}\gamma(t)\right)\dot{y}(t) + \sqrt{s}\nabla^2 f(y(t))\dot{y}(t) + \left(1 + \frac{\sqrt{s}}{2}\gamma(t)\right)\nabla f(y(t)) = 0. \tag{50}$$

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

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

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, (51) can be formulated equivalently as follows

$$\gamma_k\frac{y_{k+1} - 2y_k + y_{k-1}}{h^2} + (1-\gamma_k)\frac{y_{k+1} - y_k}{h^2} + \nabla f(y_k) + \gamma_k(\nabla f(y_k) - \nabla f(y_{k-1})) = 0. \tag{52}$$

After dividing by γ_k, we get

$$\frac{y_{k+1} - 2y_k + y_{k-1}}{h^2} + \frac{1-\gamma_k}{h\gamma_k}\frac{y_{k+1} - y_k}{h} + \frac{1}{\gamma_k}\nabla f(y_k) + (\nabla f(y_k) - \nabla f(y_{k-1})) = 0. \tag{53}$$

Building on (53), 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 (53), we obtain

$$y_{k+1} = Y(\tau_{k+1}) = Y(\tau_k) + h\dot{Y}(\tau_k) + \frac{1}{2}h^2\ddot{Y}(\tau_k) + \frac{1}{6}h^3\dddot{Y}(\tau_k) + O(h^4), \tag{54}$$

$$y_{k-1} = Y(\tau_{k-1}) = Y(\tau_k) - h\dot{Y}(\tau_k) + \frac{1}{2}h^2\ddot{Y}(\tau_k) - \frac{1}{6}h^3\dddot{Y}(\tau_k) + O(h^4). \tag{55}$$

By adding (54) and (55) we obtain

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

Moreover, (54) gives

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

By Taylor expansion of ∇f we have

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

Plugging all of the above results into (53), we obtain

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

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

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

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

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

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

$$\ddot{Y}(\tau_k)+\frac{2(1-\gamma_k)}{h(1+\gamma_k)}\dot{Y}(\tau_k)+\frac{2}{1+\gamma_k}\nabla f(Y(\tau_k))+\frac{2h\gamma_k}{1+\gamma_k}\nabla^2 f(Y(\tau_k))\dot{Y}(\tau_k)+O(h^2)=0. \tag{56}$$

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

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

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 [9] and [42]. 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 (56), 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. Assran, M., Rabbat, M.: On the convergence of Nesterov's accelerated gradient method in stochastic settings. In: The 37th International Conference on Machine Learning (ICML), vol. 119, pp. 410–420. PMLR (2020)
3. 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`
4. 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`

5. 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
6. 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
7. 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
8. 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
9. 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
10. 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
11. 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
12. 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
13. Can, B., Gurbuzbalaban, M., Zhu, L.: Accelerated linear convergence of stochastic momentum methods in Wasserstein distances. In: The 36th International Conference on Machine Learning (ICML). PMLR (2019). URL https://arxiv.org/abs/1901.07445
14. 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
15. Flammarion, N., Bach, F.: From averaging to acceleration, there is only a step-size. In: The 28th Conference on Learning Theory (COLT), vol. 40, pp. 658–695. PMLR (2015). URL https://proceedings.mlr.press/v40/Flammarion15.html
16. 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
17. Gelfand, I., Tsetlin, M.: Printszip nelokalnogo poiska v sistemah avtomatich. Optimizatsii, Dokl. AN SSSR **137**, 295–298 (1961). (in Russian)
18. Ghadimi, S., Lan, G.: Optimal stochastic approximation algorithms for strongly convex stochastic composite optimization i: A generic algorithmic framework. SIAM J. Optim. **22**(4), 1469–1492 (2012). DOI 10.1137/110848864. URL https://doi.org/10.1137/110848864
19. Ghadimi, S., Lan, G.: Optimal stochastic approximation algorithms for strongly convex stochastic composite optimization, ii: Shrinking procedures and optimal algorithms. SIAM J. Optim. **23**(4), 2061–2089 (2013). DOI 10.1137/110848876. URL https://doi.org/10.1137/110848876
20. Gupta, K., Siegel, J.W., Wojtowytsch, S.: Nesterov acceleration despite very noisy gradients. In: The International Conference on Neural Information Processing Systems (NeurIPS) (2024). URL https://arxiv.org/abs/2302.05515
21. 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
22. Kidambi, R., Netrapalli, P., Jain, P., Kakade, S.: On the insufficiency of existing momentum schemes for stochastic optimization. In: Information Theory and Applications Workshop (ITA), pp. 1–9. IEEE (2018)
23. 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
24. Laborde, M., Oberman, A.: A Lyapunov analysis for accelerated gradient methods: from deterministic to stochastic case. In: The Twenty Third International Conference on Artificial Intelligence and Statistics (AISTATS), vol. 108, pp. 602–612. PMLR (2020). URL https://proceedings.mlr.press/v108/laborde20a.html
25. Lan, G.: An optimal method for stochastic composite optimization. Math. Program. **133**(1), 365–397 (2012). DOI 10.1007/s10107-010-0434-y. URL https://doi.org/10.1007/s10107-010-0434-y
26. Lan, G.: First-order and Stochastic Optimization Methods for Machine Learning. Springer Series in the Data Sciences. Springer, Cham (2020). DOI 10.1007/978-3-030-39568-1. URL https://doi.org/10.1007/978-3-030-39568-1

27. 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)
28. Liu, J., Yuan, Y.: On almost sure convergence rates of stochastic gradient methods. In: The 35th Conference on Learning Theory (COLT), vol. 178, pp. 2963–2983. PMLR (2022). URL https://proceedings.mlr.press/v178/liu22d.html
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. Maulen-Soto, R., Fadili, J., Attouch, H., Ochs, P.: An SDE perspective on stochastic inertial gradient dynamics with time-dependent viscosity and hessian-driven damping. Optimization (2025)
31. Maulen-Soto, R., Fadili, J., Attouch, H., Ochs, P.: Stochastic inertial dynamics via time scaling and averaging. Stochastic Systems (2025)
32. 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
33. 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)
34. Orvieto, A., Kohler, J., Lucchi, A.: The role of memory in stochastic optimization. In: the 35th Conference on Uncertainty in Artificial Intelligence (UAI) (2020). URL https://arxiv.org/abs/1907.01678
35. 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
36. Polyak, B.T.: Some methods of speeding up the convergence of iterative methods. Ž. Vyčisl. Mat i Mat. Fiz. **4**, 791–803 (1964)
37. Polyak, B.T.: Introduction to Optimization. Translations Series in Mathematics and Engineering. Optimization Software, Inc., Publications Division, New York (1987)
38. Polyak, B.T.: Accelerated gradient methods revisited. In: Workshop on Variational Analysis and Applications. Erice, Italy (August 28-September 5, 2018)
39. Robbins, H., Siegmund, D.: A convergence theorem for non negative supermartingales and some applications. Optimizing Methods in Statistics pp. 233–257 (1971). URL https://doi.org/10.1016/B978-0-12-604550-5.50015-8
40. Schmidt, M., Roux, N.L., Bach, F.: Convergence rates of inexact proximal-gradient methods for convex optimization. In: The 24th International Conference on Neural Information Processing Systems (NeurIPS), pp. 1458–1466 (2011)
41. Sebbouh, O., Gower, R.M., Defazio, A.: Almost sure convergence rates for stochastic gradient descent and stochastic heavy ball. In: The 34th Conference on Learning Theory (COLT), vol. 134, pp. 3935–3971. PMLR (2021). URL https://proceedings.mlr.press/v134/sebbouh21a.html
42. 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
43. 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)
44. Sutskever, I., Martens, J., Dahl, G., Hinton, G.: On the importance of initialization and momentum in deep learning. In: The 30th International Conference on Machine Learning (ICML), vol. 28, pp. 1139–1147. PMLR, Atlanta, Georgia, USA (2013). URL https://proceedings.mlr.press/v28/sutskever13.html
45. 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
46. Yang, T., Lin, Q., Li, Z.: Unified convergence analysis of stochastic momentum methods for convex and non-convex optimization. In: The 27th International Joint Conference on Artificial Intelligence (IJCAI) (2018). URL https://arxiv.org/abs/1604.03257