Improving Stochastic Cubic Newton with Momentum

El Mahdi Chayti   **Nikita Doikov**   **Martin Jaggi**
Machine Learning and Optimization Laboratory (MLO), EPFL

Abstract

We study stochastic second-order methods for solving general non-convex optimization problems. We propose using a special version of momentum to stabilize the stochastic gradient and Hessian estimates in Newton's method. We show that momentum provably improves the variance of stochastic estimates and allows the method to converge for any noise level. Using the cubic regularization technique, we prove a global convergence rate for our method on general non-convex problems to a second-order stationary point, even when using only a single stochastic data sample per iteration. This starkly contrasts with all existing stochastic second-order methods for non-convex problems, which typically require large batches. Therefore, we are the first to demonstrate global convergence for batches of arbitrary size in the non-convex case for the Stochastic Cubic Newton. Additionally, we show improved speed on convex stochastic problems for our regularized Newton methods with momentum.

1 Introduction

1.1 Motivation

We are interested in solving general stochastic optimization problems of the form

$$\min_{\boldsymbol{x}\in\mathbb{R}^d} f(\boldsymbol{x}) := \mathbb{E}_{\xi\sim\mathcal{P}}\left[f_\xi(\boldsymbol{x})\right], \tag{1}$$

where the functions f_ξ are twice differentiable, not necessarily convex, and only reasonably have access to the distribution $\mathcal{P}$ through its samples. Problems of

this type are important in machine learning applications and can be challenging to solve due to both the *stochasticity* and *ill-conditioning* of f.

In this work, we study the class of second-order methods, which employ a *second-order stochastic oracle* of (1): for every $\boldsymbol{x} \in \mathbb{R}^d$ and random sample $\xi \in \mathcal{P}$, we assume to have access to stochastic gradient $\nabla f_\xi(\boldsymbol{x})$ and stochastic Hessian matrix $\nabla^2 f_\xi(\boldsymbol{x})$. Incorporating second-order information into an optimization scheme is known to be a powerful approach to deal with ill-conditioning of the problem (Polyak, 1987; Nocedal and Wright, 2006; Nesterov, 2018).

Indeed, for a *full-batch version* of the cubically regularized Newton's method, which uses the full gradient $\nabla f(\boldsymbol{x})$ and the full Hessian $\nabla^2 f(\boldsymbol{x})$, it is possible to establish the state-of-the-art global convergence rates to a second-order stationary point, for general non-convex functions, and *the rate is provably better* than that of first order gradient methods (Nesterov and Polyak, 2006). Thus, we can see a strong benefit in utilizing second-order information to address ill-conditioning of the problem.

However, in modern large-scale applications, it is often unfeasible to have access to a full exact oracle of f. In the presence of stochasticity, such as in machine learning applications, it becomes especially challenging to use noisy gradients and Hessians in optimization algorithms for solving (1). While there exist several approaches that use highly accurate approximations of the gradient and Hessian with subsampling over a sufficiently large batch, it remained an important open problem:

Can we design a second-order method that converges globally, for general non-convex functions, for batches as small as one sample?

1.2 Contributions

In this paper, we provide a positive answer to this question by designing an efficient stochastic second-order optimization algorithm that can work with stochastic gradients and Hessians of *arbitrarily large variance*,

Algorithm	Global complexity	Arbitrary accuracy ε	Smooth. assump.
Stochastic Gradient Descent (SGD) (Lan, 2020)	$F_0 \cdot \mathcal{O}\left(\frac{L_g}{\varepsilon^2} + \frac{L_g^{1/2}\sigma_g}{F_0^{1/2}\varepsilon^4}\right)$	yes	Lipschitz ∇f
Normalized SGD with momentum (Cutkosky and Mehta, 2020a)	$F_0 \cdot \mathcal{O}\left(\frac{L_g}{\varepsilon^2} + \frac{\sigma_g^{13/3}}{F_0^{7/3}L^{2/3}\varepsilon^{7/3}} + \frac{L^{1/2}\sigma_g^2}{\varepsilon^{7/2}}\right)$	yes	Lip. ∇f and $\nabla^2 f$
Stochastic Cubic Newton (SCN) (Chayti et al., 2024)	$F_0 \cdot \mathcal{O}\left(\frac{L^{1/2}}{\varepsilon^{3/2}} + \frac{\sigma_h}{\varepsilon^2}\right)$	only for $\varepsilon \geq \sigma_g^{3/2}$	Lipschitz $\nabla^2 f$
SCN with momentum **(ours)**	$F_0 \cdot \mathcal{O}\left(\frac{L^{1/2}}{\varepsilon^{3/2}} + \frac{L^{1/4}\sigma_h^{1/2}}{\varepsilon^{7/4}} + \frac{L^{1/2}\sigma_g^2}{\varepsilon^{7/2}}\right)$	yes	Lipschitz $\nabla^2 f$

Table 1: A comparison of first-order and second-order methods for non-convex optimization that use a single stochastic sample of gradient and Hessian per iteration. Global complexity means the number of iterations (single data samples) required to obtain an $\varepsilon > 0$ accuracy for the expected gradient norm: $\|\nabla f(\bar{\boldsymbol{x}})\| \leq \varepsilon$. $F_0 := f(\boldsymbol{x}_0) - f^\star$ is the initial functional residual, L_g denotes the Lipschitz constant of the gradient, and L denotes the Lipschitz constant of the Hessian, σ_g denotes the variance of gradient samples and σ_H stands for the variance of the Hessian samples (see precise definitions in Section 2.1). Note that L is not affected by adding to our objective f an arbitrary quadratic function, while L_g, used in first-order methods, can become arbitrarily large.

converging to a second-order stationary point for non-convex objectives. The main idea behind our algorithm is to use a special form of *momentum* for stochastic estimates of the gradient and Hessian. Thus, at iteration $t \geq 1$, we use a recursively defined vector

$$\boldsymbol{g}_t = (1-\alpha)\boldsymbol{g}_{t-1} + \alpha\nabla f_{\xi_t}\left(\boldsymbol{x}_t + \tfrac{1-\alpha}{\alpha}(\boldsymbol{x}_t - \boldsymbol{x}_{t-1})\right) \quad (2)$$

and matrix

$$\boldsymbol{H}_t = (1-\beta)\boldsymbol{H}_{t-1} + \beta\nabla^2 f_{\xi_t}(\boldsymbol{x}_t), \quad (3)$$

where $(\boldsymbol{x}_t)_{t\geq 0}$ is a sequence of points generated by the method, $0 < \alpha, \beta \leq 1$ are momentum parameters, and ξ_t is a random sample of data. We start these updates with $\boldsymbol{g}_0 = \nabla f_{\xi_0}(\boldsymbol{x}_0)$ and $\boldsymbol{H}_0 = \nabla f_{\xi_0}(\boldsymbol{x}_0)$ for some initial starting point. Then, to compute the next point $\boldsymbol{x}_{t+1}$, we perform the Cubic Newton step from the previous one using freshly estimated $\boldsymbol{g}_t$ and $\boldsymbol{H}_t$. Therefore, in a non-degenerate case, one step of our second-order method can be represented in the following standard form:

$$\boldsymbol{x}_{t+1} = \boldsymbol{x}_t - \left(\boldsymbol{H}_t + \gamma_t\boldsymbol{I}\right)^{-1}\boldsymbol{g}_t, \quad (4)$$

where $\gamma_t > 0$ is a parameter that comes from the cubic regularization of the second-order model, and which ensures that $\boldsymbol{H}_t + \gamma_t\boldsymbol{I} \succeq 0$ (see Algorithm 1 in Section 3 for a formal description of our method).

Note that if we choose $\alpha = 1$ and $\beta = 1$ in our update of $\boldsymbol{g}_t$ and $\boldsymbol{H}_t$, we obtain simply $\boldsymbol{g}_t = \nabla f_{\xi_t}(\boldsymbol{x}_t)$ and $\boldsymbol{H} = \nabla^2 f_{\xi_t}(\boldsymbol{x}_t)$ which corresponds to a straightforward 'naive' implementation of the stochastic Cubic Newton. In our work, we show that in the presence of noise, it is essential to pick $\alpha, \beta < 1$, which adds momentum to the updates and which ensures convergence of the method for samples with arbitrarily large variance.

We summarize our contributions as follows.

- We propose combining the special form of momentum (2), (3) with stochastic second-order methods and demonstrate how it can reduce the variance of stochastic samples by simulating large batches using past estimates of the gradient and Hessian.
- We prove that the Stochastic Cubic Newton method with momentum enjoys an improved convergence rate, for general non-convex problems (see Table 1). In particular, we show that the method converges for any choice of batch size, *even just a single sample*. Importantly, our result is obtained under the standard assumptions commonly used in previous literature.
- We also show an improved convergence rate for our method in the convex case (see Section 3.3), again for any sample size. Notably, we prove this rate for the primal Cubic Newton scheme, without the need to use additional extragradient steps.
- Illustrative numerical experiments are provided.

1.3 Related work

When the objective f is non-convex, problem (1) is especially complicated since finding a global minimum is NP-hard in general (Hillar and Lim, 2013). Hence, the reasonable goal is to look for approximate solutions. *First-order gradient methods* (Cauchy, 1847; Robbins and Monro, 1951; Kiefer and Wolfowitz, 1952; Lan, 2020) only use first-order information (gradients) and can, generally, only guarantee convergence to approximate stationary points i.e., a point with a small gradient norm. For non-convex functions, a stationary point may be a saddle point or even a local maximum, which might not be desirable. To obtain faster convergence rates for ill-conditioned problems and ensure a second-order stationary condition for an approximate

stationary point, we can use second-order information.

Second-order optimization. Among many variants of the Newton algorithm, one of the most theoretically established is the Cubic Newton method (Griewank, 1981; Nesterov and Polyak, 2006; Cartis et al., 2011a,b), which enjoys many favorable properties such as faster global convergence (compared to first-order methods), and local quadratic convergence, as for the classical Newton's method (Fine, 1916; Bennett, 1916; Kantorovich, 1948; Polyak, 2007).

Inexact versions of the Cubic Newton with approximated gradients and Hessians studied intensively in the literature (Cartis et al., 2011a,b; Wang et al., 2018), including *low-rank quasi-Newton approximations* (Benson and Shanno, 2018; Kamzolov et al., 2023; Scieur, 2024; Doikov et al., 2024), *stochastic subspace methods* (Hanzely et al., 2020; Zhao et al., 2024), and *finite-difference approximations* (Cartis et al., 2012; Grapiglia et al., 2022; Doikov and Grapiglia, 2023). However, most of these works assume access to deterministic (full batch) oracle information about the objective function, which can be very expensive or even unfeasible in the case of large datasets.

The main approach to overcome this issue consists in using *stochastic* gradient and Hessian estimates *subsampled* from the data (Xu et al., 2016; Kohler and Lucchi, 2017; Xu et al., 2017; Nilesh et al., 2018; Ghadimi et al., 2017; Cartis and Scheinberg, 2018; Doikov and Nesterov, 2020; Agafonov et al., 2020). In the subsampling technique, we rely on the use of *sufficiently large batches* of data to reduce the effects of noise on the method's convergence, thereby making the gradient and Hessian estimates highly accurate. However, as was also noticed by (Chayti et al., 2024), this technique *does not guarantee convergence* of the Cubic Newton for small batches. This is in contrast to stochastic first-order methods, that often possess convergence, starting from an arbitrary initialization and even for one sample per iteration (Lan, 2020).

In fact, all existing analyses of the subsampled stochastic Cubic Newton method can guarantee convergence only to a ball of radius $\mathcal{O}(\sigma_g^{3/2})$, where σ_g is the variance of the gradient noise defined in the usual manner (see Table 1). Therefore, if the noise level is large, the method can't guarantee convergence to an arbitrary level of accuracy ε.

Convex optimization. This problem seems to be widely present even in the convex case. For example, (Agafonov et al., 2020) shows convergence in terms of the global functional residual up to a ball $\mathcal{O}(\sigma_g D)$ for stochastic high-order methods. At the same time, for convex problems, it is possible to incorporate stochastic Newton steps into *extragradient* (Korpelevich, 1976; Ne-

mirovski, 2004) and *accelerated* (Nesterov, 1983, 2008) schemes, which provides global convergence guarantees for arbitrary sample size (Antonakopoulos et al., 2022; Agafonov et al., 2024). However, extending this approach to non-convex functions is challenging even for first-order methods.

In our work, we are mainly focused on non-convex functions; thus we propose a different strategy that consists of reusing past gradient and Hessian estimates to simulate a large batch without ever having to actually sample one.

Momentum. The idea of using momentum has been extensively explored for first-order optimization. It was initially proposed by (Polyak, 1964) in his Heavy-Ball method, aimed at improving the convergence properties of classical gradient descent. It was shown in (Polyak, 1964) (see also (Polyak, 1987)) that the Heavy-Ball method achieves the accelerated optimal convergence rate for deterministic minimization of convex quadratic functions. Recent results show the benefits of using momentum for deterministic optimization, taking into account the intrinsic distribution of the Hessian spectrum (Scieur and Pedregosa, 2020), Hölder continuity of the Hessian (Marumo and Takeda, 2024), and also demonstrate the impossibility of acceleration for general strongly convex functions with Lipschitz gradient (Goujaud et al., 2023).

In stochastic optimization, the use of momentum in SGD and its variants such as Adam (Kingma and Ba, 2017) is very popular. Different forms of momentum were established in (Tran and Cutkosky, 2022; Cutkosky and Orabona, 2020; Arnold et al., 2019) that lead to improved convergence rates. Surprisingly, this simple, yet effective idea, has not been explored for stochastic second-order methods, in particular, the Cubic Newton. The only works we were able to find combining both momentum and the Cubic Newton are (Wang et al., 2019; Gao and Ng, 2022), and they both consider a different version of momentum than ours. Most importantly, they consider deterministic optimization without stochasticity.

Reusing past samples is a very well-used idea in machine learning to deal with stochasticity and smooth out estimates (mostly of gradients as this was mainly used for first-order methods). It is very easy to see that momentum, in this sense, introduces bias to the estimates (since the present is not necessarily the same as the past), quantifying this bias is necessary for analyzing any type of momentum. The variants of momentum in (Tran and Cutkosky, 2022; Cutkosky and Orabona, 2020; Arnold et al., 2019), are but different ways to algorithmically reduce this bias (under different assumptions). Extending the analyses of momentum to

Cubic Newton is particularly challenging as there is no closed-form solution of the update at each step, rendering controlling the bias incurred from using momentum very complicated.

2 Preliminaries

Notation. We denote by $\|\boldsymbol{x}\| := \langle \boldsymbol{x}, \boldsymbol{x}\rangle^{1/2}$, $\boldsymbol{x} \in \mathbb{R}^d$, the standard Euclidean norm for vectors, and the spectral norm for symmetric matrices by $\|\boldsymbol{H}\| := \max\{\lambda_{\max}(\boldsymbol{H}), -\lambda_{\min}(\boldsymbol{H})\}$, where $\boldsymbol{H} = \boldsymbol{H}^\top \in \mathbb{R}^{d \times d}$. For functions h_1, h_2, we write $h_1 = \mathcal{O}(h_2)$ to denote $h_1 \leq Ch_2$ for some universal numerical constant C, which does not depend on the problem class parameters.

As is customary for second-order methods, we assume the objective function f has L-Lipschitz Hessians, which is, for all $\boldsymbol{x}, \boldsymbol{y} \in \mathbb{R}^d$:

$$\|\nabla^2 f(\boldsymbol{x}) - \nabla^2 f(\boldsymbol{y})\| \leq L\|\boldsymbol{x} - \boldsymbol{y}\|. \tag{5}$$

We also assume that the objective is lower bounded from below, $f^\star := \min_{\boldsymbol{x}} f(\boldsymbol{x}) > -\infty$, and denote $F_0 = f(\boldsymbol{x}_0) - f^\star$ for some initialization of the state $\boldsymbol{x}_0$ that we fix throughout this paper without loss of generality.

The (Stochastic) Cubic Newton algorithm can be written as the following update rule of the current state $\boldsymbol{x}_t$:

$$\boldsymbol{x}_{t+1} \in \arg\min_{\boldsymbol{y} \in \mathbb{R}^d} \Big\{ \Omega_{M,\boldsymbol{g},\boldsymbol{H}}(\boldsymbol{y}, \boldsymbol{x}_t) \Big\}, \tag{6}$$

where Ω is the cubically regularized quadratic model of the objective, defined by

$$\Omega_{M,\boldsymbol{g},\boldsymbol{H}}(\boldsymbol{y}, \boldsymbol{x}_t) := \langle \boldsymbol{g}_t, \boldsymbol{y} - \boldsymbol{x}_t\rangle$$
$$+ \frac{1}{2}\langle \boldsymbol{H}_t(\boldsymbol{y} - \boldsymbol{x}_t), \boldsymbol{y} - \boldsymbol{x}_t\rangle + \frac{M}{6}\|\boldsymbol{y} - \boldsymbol{x}_t\|^3, \tag{7}$$

with regularization parameter $M > 0$, which is typically greater than the Lipschitz constant L, whereas $\boldsymbol{g}_t$ and $\boldsymbol{H}_t$ are (stochastic) estimates of the gradient and Hessian at $\boldsymbol{x}_t$. Note that subproblem in (6) is generally non-convex, when $\boldsymbol{H}_t \not\succeq 0$, and can have several isolated global minima. However, it is possible to solve this problem efficiently with the same arithmetic cost as for the classical Newton step. In a non-degenerate case, one iteration of the form (6) can be represented in the standard form (4), regularizing the Hessian by identity matrix with a certain regularization coefficient γ_t, which can be found by solving the corresponding univariate dual problem (see details in (Conn et al., 2000; Nesterov and Polyak, 2006; Cartis et al., 2011a)).

Examples of choices for $\boldsymbol{g}_t$ and $\boldsymbol{H}_t$:

1. Deterministic case: $\boldsymbol{g}_t = \nabla f(\boldsymbol{x}_t)$, $\boldsymbol{H}_t = \nabla^2 f(\boldsymbol{x}_t)$.

2. Stochastic case, subsampled approaches: we sample batches $\mathcal{B}_g, \mathcal{B}_h$ of sizes b_g, b_h respectively and set

$$\begin{aligned} \boldsymbol{g}_t &= \frac{1}{b_g} \sum_{i \in \mathcal{B}_g} \nabla f_i(\boldsymbol{x}_t), \\ \boldsymbol{H}_t &= \frac{1}{b_h} \sum_{i \in \mathcal{B}_h} \nabla^2 f_i(\boldsymbol{x}_t). \end{aligned} \tag{8}$$

In the deterministic case, Cubic Newton can be understood as replacing the function f by its second-order Taylor approximation augmented by cubic regularization to control the region where this approximation is valid. When using stochastic samples, we further replace the exact Taylor model with the stochastic (7).

A point $\boldsymbol{x}$ is denoted a (ε, c)-*approximate second-order local minimum* if it satisfies:

$$\|\nabla f(\boldsymbol{x})\| \leq \varepsilon \quad \text{and} \quad \lambda_{min}(\nabla^2 f(\boldsymbol{x})) \geq -c\sqrt{\varepsilon},$$

where $\varepsilon, c > 0$ are given tolerance parameters. Let us define the following accuracy measure (see Nesterov and Polyak (2006)):

$$\mu_c(\boldsymbol{x}) := \max\Big(\|\nabla f(\boldsymbol{x})\|^{3/2}, \frac{-\lambda_{min}(\nabla^2 f(\boldsymbol{x}))^3}{c^{3/2}}\Big).$$

Note that this definition implies that if $\mu_c(\boldsymbol{x}) \leq \varepsilon^{3/2}$ then $\boldsymbol{x}$ is an (ε, c)-approximate *local minimum*.

2.1 Stochastic Cubic Newton

For the sake of brevity, let us denote $\boldsymbol{\varepsilon}_t = \boldsymbol{g}_t - \nabla f(\boldsymbol{x}_t)$ (resp. $\boldsymbol{\Sigma}_t = \boldsymbol{H}_t - \nabla^2 f(\boldsymbol{x}_t)$) the gradient (resp. Hessian) approximation error and $r_t = \|\boldsymbol{x}_{t+1} - \boldsymbol{x}_t\|$ the norm of the state update.

The main standard Lemma that we use for the analysis of one step of (6) is the following:

> **Lemma 2.1.** *For any $M \geq L$, the progress made by the update rule* (6) *satisfies:*
>
> $$f(\boldsymbol{x}_t) - f(\boldsymbol{x}_{t+1}) \geq \mathcal{O}(1)\Big(\frac{1}{\sqrt{M}}\mu_M(\boldsymbol{x}_{t+1})$$
> $$+ Mr_t^3 - \frac{\|\boldsymbol{\varepsilon}_t\|^{3/2}}{\sqrt{M}} - \frac{\|\boldsymbol{\Sigma}_t\|^3}{M^2}\Big).$$

In light of Lemma 2.1, we see that we need to control the $(3/2)$−th moment of the gradient noise $\|\boldsymbol{\varepsilon}_t\|^{3/2}$ and the third moment of the Hessian noise $\|\boldsymbol{\Sigma}_t\|^3$, this is drastically different from the second moment usually encountered for first-order methods. The main complication that makes transferring methods from the first-order world to the second-order world is that we do not have an explicit expression of (6), making it very difficult to deal with the noise. In fact, Lemma 2.1 gives the same bound if we assume the noise is deterministic (which is typically worse) or randomly centered, this

suggests either a gap in theory or an intrinsic behavior of the method. It remains an open question to establish stochastic lower bounds for second-order methods that enable distinguishing between these cases.

Noise assumptions. the usual noise assumptions made in the literature (Kohler and Lucchi, 2017; Xu et al., 2017; Tripuraneni et al., 2017; Agafonov et al., 2024) are the following:

Assumption 2.2 (gradient noise). *All samples ξ are sampled independently from the past, and there exists $\sigma_g < \infty$ such that*

$$\mathbb{E}\nabla f_\xi(\boldsymbol{x}) = \nabla f(\boldsymbol{x}) \quad and$$

$$\mathbb{E}\|\nabla f_\xi(\boldsymbol{x}) - \nabla f(\boldsymbol{x})\|^2 \leq \sigma_g^2 .$$

Note that by using Jensen's inequality, we have: $\mathbb{E}\|\nabla f_\xi(\boldsymbol{x}) - \nabla f(\boldsymbol{x})\|^{3/2} \leq \sigma_g^{3/2}$. By independence, the subsampled gradient in (8) satisfies:

$$\mathbb{E}\|\boldsymbol{\varepsilon}_t\|^{3/2} \leq \frac{\sigma_g^{3/2}}{b_g^{3/4}},$$

where b_g is the batch size of the gradient samples.

Controlling the Hessian noise is much more complicated since we cannot use Jensen's inequality anymore, and it is easy to construct distributions with finite second-moment and infinite third moment. Thus we naturally need to make stronger assumptions.

Assumption 2.3 (Hessian noise). *All samples ξ are sampled independently from the past, and there exist positive constants σ_h, δ_h such that*

$$\mathbb{E}\nabla^2 f_\xi(\boldsymbol{x}) = \nabla^2 f(\boldsymbol{x}),$$

$$\mathbb{E}\|\nabla^2 f_\xi(\boldsymbol{x}) - \nabla^2 f(\boldsymbol{x})\|^2 \leq \sigma_h^2, \quad and$$

$$\|\nabla^2 f_\xi(\boldsymbol{x}) - \nabla^2 f(\boldsymbol{x})\| \leq \delta_h \quad (a.s).$$

Assumption 2.3, coupled with Theorem A.1 from (Chen et al., 2012) leads for example to the following bound on the subsampled Hessian in (8):

$$\mathbb{E}\|\boldsymbol{\Sigma}_t\|^3 = \mathcal{O}\Big(\log(d)^{3/2}\frac{\sigma_h^3}{b_h^{3/2}} + \log(d)^3\frac{\delta_h^3}{b_h^3}\Big)$$

$$\underset{b_h \gg 1}{\approx} \mathcal{O}\Big(\frac{\sigma_h^3}{b_h^{3/2}}\Big),$$

where b_h is the batch size of Hessian samples.

Under these assumptions, as shown by (Chayti et al., 2024), using the estimates in (8), we get that for T steps of (6) and for $\boldsymbol{x}_{out}^T$ sampled uniformly at random from the iterates $(\boldsymbol{x}_t)_{1\leq t\leq T}$:

$$\mathbb{E}\mu_M(\boldsymbol{x}_{out}^T) = \mathcal{O}\left(\frac{\sqrt{M}F_0}{T} + \frac{\sigma_h^3}{M^{3/2}b_h^{3/2}} + \frac{\sigma_g^{3/2}}{b_g^{3/4}}\right). \quad (9)$$

What is noteworthy about (9) is that for $b_g = b_h = 1$, even with the best choice of M, (9) can only guarantee convergence $\mathbb{E}\mu_M(\boldsymbol{x}_{out}^T) \leq \varepsilon$ up to $\varepsilon \geq \mathcal{O}(\sigma_g^{3/2})$.

2.2 Momentum to the rescue

The main idea of this paper is to reuse past iterates to "forge" a large batch. This concept is not new in the first-order optimization literature; The stochastic momentum or Polyak's momentum (Polyak, 1964), takes the form of an exponential moving average:

$$\boldsymbol{g}_t = (1 - \alpha_t)\boldsymbol{g}_{t-1} + \alpha_t \nabla f_{\xi_t}(\boldsymbol{x}_t) \quad \text{(HB)}$$

for some $\alpha_t \in (0, 1]$. This momentum is known to suffer from a high bias. To understand this, imagine $\boldsymbol{g}_{t-1} = \nabla f(\boldsymbol{x}_{t-1})$. Then, even in this perfect case,

$$\mathbb{E}\big[\boldsymbol{g}_t - \nabla f(\boldsymbol{x}_t)\big] = (1 - \alpha_t)(\nabla f(\boldsymbol{x}_{t-1}) - \nabla f(\boldsymbol{x}_t)) \neq \boldsymbol{0}.$$

Besides, its theoretical analysis usually involves assuming that the gradient of f is Lipschitz, an assumption that typically slows down the Cubic Newton method, and which is not required in our work (see also the appendix, where we additionally discuss using the classical momentum in the Cubic Newton).

At least three corrections were proposed for (HB), the MVR (for Momentum based Variance Reduction) momentum (Cutkosky and Orabona, 2020), the IT (stands for Implicit "Gradient" Transport) momentum (Arnold et al., 2019) and the second-order momentum (SOM) Tran and Cutkosky (2022).

To describe these corrections, let us consider that we observed a sequence of samples $(\xi_j)_{j\geq 0}$ and its corresponding estimates $(\pi_{\xi_j}(\boldsymbol{x}_j))_{j\geq 0}$, which can be either sampled gradients or Hessians. Let $\alpha_t := 1/t$ and define the following quantity: $\Pi_t(\boldsymbol{x}) := \frac{1}{t}\sum_{j=0}^t \pi_{\xi_j}(\boldsymbol{x})$.

For illustration, let us assume that the functions π_ξ are all affine, $\pi_\xi(\boldsymbol{x}) = \boldsymbol{A}\boldsymbol{x} + \boldsymbol{b}_\xi$, with the same matrix $\boldsymbol{A}$ and different bias parts $\boldsymbol{b}_\xi$ (e.g. π_ξ are the gradients of a quadratic function). Then we can easily show that

$$\Pi_t(\boldsymbol{x}_t) = (1 - \alpha_t)\Big[\Pi_{t-1}(\boldsymbol{x}_{t-1}) + \nabla\pi_{\xi_t}(\boldsymbol{x}_t)(\boldsymbol{x}_t - \boldsymbol{x}_{t-1})\Big]$$

$$+ \alpha_t\pi_{\xi_t}(\boldsymbol{x}_t) \quad \text{(SOM)}$$

$$= (1 - \alpha_t)\Big[\Pi_{t-1}(\boldsymbol{x}_{t-1}) + \pi_{\xi_t}(\boldsymbol{x}_t) - \pi_{\xi_t}(\boldsymbol{x}_{t-1})\Big]$$

$$+ \alpha_t\pi_{\xi_t}(\boldsymbol{x}_t) \quad \text{(MVR)}$$

$$= (1 - \alpha_t)\Pi_{t-1}(\boldsymbol{x}_{t-1})$$

$$+ \alpha_t\pi_{\xi_t}\Big(\boldsymbol{x}_t + \frac{1 - \alpha_t}{\alpha_t}(\boldsymbol{x}_t - \boldsymbol{x}_{t-1})\Big). \quad \text{(IT)}$$

To have the corresponding expressions for the gradient (resp. Hessian) it suffices to take $\pi = \nabla f$ (resp. $\pi = \nabla^2 f$) and replace $\Pi_t(\boldsymbol{x}_t)$ by $\boldsymbol{g}_t$ (resp. $\boldsymbol{H}_t$).

Although the three expressions we show are equivalent for affine π, they generally behave differently and need different assumptions to analyze their behavior.

Many combinations of momentum for the gradient and Hessian are possible but we will only consider in the main paper, (IT) for the gradient and (HB) for the Hessian as this is the only combination that led to an improvement and most importantly to convergence for arbitrary batch sizes for the stochastic Cubic Newton without needing any additional assumption. We consider other combinations in the appendix.

3 Main theoretical analysis

Based on the previous discussion, we consider the following update rules for the gradient and Hessian estimates:

$$
\begin{aligned}
\boldsymbol{g}_t &= (1 - \alpha_t)\boldsymbol{g}_{t-1} \\
&+ \alpha_t \nabla f_{\xi_t}\big(\boldsymbol{x}_t + \tfrac{1-\alpha_t}{\alpha_t}(\boldsymbol{x}_t - \boldsymbol{x}_{t-1})\big),
\end{aligned} \tag{10}
$$

and

$$
\boldsymbol{H}_t = (1 - \beta_t)\boldsymbol{H}_{t-1} + \beta_t \nabla^2 f_{\xi_t}(\boldsymbol{x}_t), \tag{11}
$$

where $\alpha_t, \beta_t \in (0, 1]$. The (IT) momentum in (10), is very similar to the Heavy Ball (HB) momentum and only differs in where the new gradient is evaluated. (HB) evaluates the new gradient at the current state $\boldsymbol{x}_t$, while (IT) evaluates it at a "transported" location: $\boldsymbol{y}_t = \boldsymbol{x}_t + \tfrac{1-\alpha_t}{\alpha_t}(\boldsymbol{x}_t - \boldsymbol{x}_{t-1})$, thus these two moments have the same cost but (IT) is typically better.

Consequently, we consider the following algorithm:

Algorithm 1 Cubic Newton with momentum

Require: $\boldsymbol{x}_0 \in \mathbb{R}^d$, $M_t > 0$, $\alpha_t, \beta_t \in (0, 1]$, T.
1: **for** $t = 0, \ldots, T - 1$ **do**
2:     Sample ξ_t^g, ξ_t^h.   # can be equal
3:     **if** $t = 0$ **then**
4:         Compute $\boldsymbol{g}_t = \nabla_{\xi_t^g} f(\boldsymbol{x}_t), \boldsymbol{H}_t = \nabla_{\xi_t^h}^2 f(\boldsymbol{x}_t)$.
5:     **else**
6:         Compute $\boldsymbol{g}_t, \boldsymbol{H}_t$ using (10), (11).
7:     Take cubic step
8:     $\boldsymbol{x}_{t+1} \in \arg\min_{\boldsymbol{y} \in \mathbb{R}^d} \Omega_{M_t, \boldsymbol{g}_t, \boldsymbol{H}_t}(\boldsymbol{y}, \boldsymbol{x}_t)$
    return $\boldsymbol{x}_{out}^T$ uniformly at random from $(\boldsymbol{x}_i)_{0 \leq i \leq T}$

To keep the analysis of Algorithm 1 simple, we consider our hyperparameters constant: the two "momentum parameters" $\alpha_t = \alpha, \beta_t = \beta$ and the regularization parameter $M_t = M$.

Note about noise Assumptions 2.2,2.3. We will assume we can bound the noise at the initial state $\boldsymbol{x}_0$ slightly differently, and denote $\sigma_{g,0} \leq \sigma_g, \sigma_{h,0} \leq \sigma_h$ its corresponding gradient and Hessian noise levels. We do

this because, in some cases of our analysis (especially the convex case), the initial noise can give rise to slower terms in our convergence rates, such terms can be easily dealt with by making sure the initial gradient and Hessian estimates are of good quality (smaller variance), which can be done using large batches but only for the initial state. However, we stress that we do not need such a trick and can always guarantee convergence for any batch size even without it.

3.1 Momentum bounds for Cubic Newton

Gradient momentum bound. Based on the form of update (10), we can write its error $\boldsymbol{\varepsilon}_t$ recursively as:

$$
\boldsymbol{\varepsilon}_t = (1-\alpha)\boldsymbol{\varepsilon}_{t-1} + (1-\alpha)\boldsymbol{z}(\boldsymbol{x}_{t-1}, \boldsymbol{x}_t) + \alpha\boldsymbol{z}(\boldsymbol{y}_t, \boldsymbol{x}_t) + \alpha\boldsymbol{n}_t,
$$

where we define

$$
\begin{aligned}
\boldsymbol{n}_t &:= \nabla f_{\xi_t}(\boldsymbol{y}_t) - \nabla f(\boldsymbol{y}_t) \quad \text{and} \\
\boldsymbol{z}(\boldsymbol{a}, \boldsymbol{b}) &:= \nabla f(\boldsymbol{a}) - \nabla f(\boldsymbol{b}) - \nabla^2 f(\boldsymbol{b})(\boldsymbol{a} - \boldsymbol{b}).
\end{aligned}
$$

Note that (5) implies $\|\boldsymbol{z}(\boldsymbol{a}, \boldsymbol{b})\| \leq \frac{L}{2}\|\boldsymbol{a} - \boldsymbol{b}\|^2$.

Lemma 3.1. *For the gradient update defined in* (10)*, we show that the sequence* $(\boldsymbol{\varepsilon}_t)_{t \geq 0}$ *of gradient errors associated to it satisfies:*

$$
\frac{1}{T}\sum_{t=0}^{T-1} \mathbb{E}\|\boldsymbol{\varepsilon}_t\|^{3/2} = \mathcal{O}\Big(\overbrace{\alpha^{3/4}\sigma_g^{3/2}}^{variance}
$$

$$
+ \underbrace{\frac{\sigma_{g,0}^{3/2}}{\alpha T} + \frac{(1-\alpha)^{3/2}L^{3/2}}{\alpha^3}\frac{1}{T}\sum_{t=0}^{T-1}\mathbb{E}r_t^3}_{bias}\Big).
$$

Lemma 3.1 shows that the average error of (10) can be decomposed into two terms: **(variance)** term $\alpha^{3/4}\sigma_g^{3/2}$ and a **(bias)** $\frac{\sigma_{g,0}^{3/2}}{\alpha T} + \frac{(1-\alpha)^{3/2}L^{3/2}}{\alpha^3}\frac{1}{T}\sum_{t=0}^{T-1}\mathbb{E}r_t^3$. The **(variance)** term shows a smaller variance for a value of α that is small, while the **(bias)** term shows the opposite effect, this means that we should find a sweet spot to balance both terms. Note how the bias can also be decomposed into the **(initial bias)** $\frac{\sigma_{g,0}^{3/2}}{\alpha T}$ and the rest as **(iterate bias)** (iterates are different). The **initial bias** can be controlled using the initial gradient sample if needed. In contrast, the **iterate bias** can be controlled using the term Mr_t^3 that we left in Lemma 2.1 (the main reason why other momentums do not work, in theory, is that they lead to an iterate bias with a power of r_t different than 3).

Hessian momentum bound. Nearly the same story can be told about the Hessian momentum defined in (11). We show that its error can be written recursively as:

$$
\boldsymbol{\Sigma}_t = (1-\beta)\boldsymbol{\Sigma}_{t-1} + (1-\beta)\boldsymbol{Z}(\boldsymbol{x}_{t-1}, \boldsymbol{x}_t) + \beta\boldsymbol{N}_t,
$$

where

$$\boldsymbol{N}_t := \nabla^2 f_{\xi_t}(\boldsymbol{y}_t) - \nabla^2 f(\boldsymbol{y}_t) \quad \text{and}$$

$$\boldsymbol{Z}(\boldsymbol{a}, \boldsymbol{b}) := \nabla^2 f(\boldsymbol{a}) - \nabla^2 f(\boldsymbol{b}).$$

Again, (5) implies that $\|\boldsymbol{Z}(\boldsymbol{a}, \boldsymbol{b})\| \leq L\|\boldsymbol{a} - \boldsymbol{b}\|$.

Lemma 3.2. *For the Hessian update defined in* (11), *we show that the sequence* $(\boldsymbol{\Sigma}_t)_{t \geq 0}$ *of Hessian errors associated to it satisfies:*

$$\frac{1}{T} \sum_{t=0}^{T-1} \mathbb{E}\|\boldsymbol{\Sigma}_t\|^3 = \mathcal{O}\left(\beta^{3/2} \tilde{\sigma}_h^3(\beta) \right.$$

$$\left. + \frac{1}{\beta T} \sigma_{h,0}^3 + \frac{(1-\beta)L^3}{\beta^3} \frac{1}{T} \sum_{t=0}^{T-1} \mathbb{E} r_t^3 \right),$$

where

$$\tilde{\sigma}_h^3(\beta) := \mathcal{O}((\log(d)^{3/2}\sigma_h^3 + \log(d)^3 \beta^{3/2}\delta_h^3)$$

$$\underset{\beta \ll 1}{\approx} \mathcal{O}(\sigma_h^3).$$

Lemma 3.2 has the same structure as Lemma 3.1 and the same remarks can be said about the bias and variance terms it contains.

To choose the momentum parameters α, β, we proceed as follows:

Controlling iterate bias terms. As mentioned before, we use the term $M r_t^3$ in Lemma 2.1 to compensate for the iterate bias terms we get in Lemmas 3.1, 3.2. We show we can get rid of this bias by choosing $\alpha, \beta \in (0, 1]$ such that:

$$\frac{4\sqrt{3}L^{3/2}}{\alpha^3} + \frac{73}{M^{3/2}} \frac{9L^3}{\beta^3} - \frac{M^{3/2}}{72} \leq 0. \quad (12)$$

To satisfy (26), it suffices to take:

$$\alpha \geq 10\sqrt{\frac{L}{M}} \quad \text{and} \quad \beta \geq 46\frac{L}{M}, \quad (13)$$

Controlling the initial bias terms. To accomplish this, we further choose

$$\alpha \in \underset{\alpha \in (0,1]}{\arg\min} \; \alpha^{3/4}\sigma_g^{3/2} + \frac{\sigma_{g,0}^{3/2}}{\alpha T} = \frac{a_g^{6/7}}{T^{4/7}}$$

and $\quad (14)$

$$\beta \in \underset{\beta \in (0,1]}{\arg\min} \; \beta^{3/2}\sigma_h^3 + \frac{1}{\beta T}\sigma_{h,0}^3 = \frac{a_h^{6/5}}{T^{2/5}},$$

where, for simplicity, we denoted $a_g = \frac{\sigma_{g,0}}{\sigma_g} \leq 1$ and $a_h = \frac{\sigma_{h,0}}{\sigma_h} \leq 1$.

All in all, based on (13) and (14), we choose

$$\alpha = \max\left(\frac{a_g^{6/7}}{T^{4/7}}, 10\sqrt{\frac{L}{M}}\right), \; \beta = \max\left(\frac{a_h^{6/5}}{T^{2/5}}, 46\frac{L}{M}\right). \quad (15)$$

To ensure $\alpha, \beta \leq 1$ it suffices to take $M \geq 100L$.

We are ready to state our main convergence results.

3.2 Non-convex rates

In the general non-convex case, we show the following theorem:

Theorem 3.3. *For any* $\alpha, \beta \in (0, 1]$ *and* $M \geq L$ *satisfying* (26), *the iterates of Algorithm 1 are such that:*

$$\mathbb{E}\mu_M(\boldsymbol{x}_{out}) = \mathcal{O}\left(\frac{\sqrt{M}F_0}{T} + \frac{\sigma_{h,0}^3}{M^{3/2}\beta T} + \frac{\sigma_{g,0}^{3/2}}{\alpha T} \right.$$

$$\left. + \frac{\beta^{3/2}\sigma_h^3}{M^{3/2}} + \alpha^{3/4}\sigma_g^{3/2} \right).$$

As a sanity check, if we take $\alpha = \beta = 1$ (meaning we do not use momentum), we get the same result (9) as past works (Chayti et al., 2024) (up to small $\mathcal{O}(1/T)$ terms). The advantage of our approach is that we can pick α, β small that can satisfy the condition (26) such as the choices in (15).

Corollary 3.4. *For* α, β *defined in* (15), *and any* $M \geq 100L$, *we have:*

$$\mathbb{E}\mu_M(\boldsymbol{x}_{out}) = \mathcal{O}\left(\frac{\sqrt{M}F_0}{T} + \frac{L^{3/2}\sigma_h^3}{M^3} + \frac{L^{3/8}}{M^{3/8}}\sigma_g^{3/2} \right.$$

$$\left. + \frac{a_g^{9/14}\sigma_g^{3/2}}{T^{3/7}} + \frac{a_h^{9/5}\sigma_h^3}{M^{3/2}T^{3/5}} \right).$$

The last two terms represent the effect of the initial gradient and Hessian noise on convergence and we can see that they decay with T.

By choosing M such that it minimizes the right-hand side in Corollary 3.4, we get:

$$\mathbb{E}\mu_M(\boldsymbol{x}_{out}) = \mathcal{O}\left(\frac{\sqrt{L}F_0}{T} \right.$$

$$+ \frac{L^{3/14}F_0^{6/7}\sigma_h^{3/7}}{T^{6/7}} + \frac{L^{3/14}F_0^{3/7}\sigma_g^{6/7}}{T^{3/7}}$$

$$\left. + \frac{a_g^{9/14}\sigma_g^{3/2}}{T^{3/7}} + \frac{a_h^{9/20}\sigma_h^{3/4}F_0^{1/2}}{T^{9/10}} \right). \quad (16)$$

The best old rate in T is $\mathcal{O}(\frac{\sqrt{L}F_0}{T} + \left(\frac{F_0\sigma_h}{T}\right)^{3/4} + \sigma_g^{3/2})$. Note that we improve both the σ_h term from $\mathcal{O}(T^{-3/4})$ to $\mathcal{O}(T^{-6/7})$ and, most importantly, we change the constant term $\mathcal{O}(\sigma_g^{3/2})$ to a decreasing term $\mathcal{O}(T^{-3/7})$, thus guaranteeing convergence for any batch size.

3.3 Extension to the convex case

Let us consider the case when the objective function f is convex[1]. We assume that all iterates $(\boldsymbol{x}_t)_{t \geq 0}$ are

---

[1] Note that we cover, in the appendix, the more general composite problems: $\min_{\boldsymbol{x} \in \mathbb{R}^d} F(\boldsymbol{x}) := \mathbb{E}_{\xi \sim \mathcal{P}}\left[f_\xi(\boldsymbol{x})\right] + \psi(\boldsymbol{x})$,

bounded within a ball of radius D: $\|\boldsymbol{x}_t - \boldsymbol{x}^\star\| \leq D$, where $\boldsymbol{x}^\star$ is any of the global minimizers that we assume to exist. This assumption is customary and can be easily achieved, for example, by introducing an additional ball constraint to our problem. These assumptions provide us with the following inequality:

$$f(\boldsymbol{x}) - f^\star \leq \langle \nabla f(\boldsymbol{x}), \boldsymbol{x} - \boldsymbol{x}^\star \rangle \leq D\|\nabla f(\boldsymbol{x})\|, \quad (17)$$

which means that convergence to a stationary point implies convergence to a global minimum. Under inequality (17), previous results (Agafonov et al., 2020; Chayti et al., 2024) give the following rate for the Stochastic Cubic Newton:

$$\mathbb{E}[f(\boldsymbol{x}_{out}^T)] - f^\star = \mathcal{O}\Big(\frac{LD^3}{T^2} + \frac{\sigma_h D^2}{T} + \sigma_g D\Big). \quad (18)$$

We observe the same problem as in the non-convex case: the method converges only to a neighborhood of the solution, and it cannot achieve accuracy $\varepsilon \leq \sigma_g D$.

Theorem 3.5. *For α, β defined in (15) and $M \geq 100L$, when the objective f is convex, the iterates of Algorithm 1 satisfy:*

$$\min_{0 \leq t \leq T-1} F_t = \mathcal{O}\Big(\frac{MD^3}{T^2} + \frac{LD\tilde{\sigma}_h^2}{M^2} + \frac{L^{1/4}}{M^{1/4}}D\sigma_g$$

$$+ F_0 e^{-T/3} + \frac{a_g^{3/7}D\sigma_g}{T^{2/7}} + \frac{a_h^{6/5}D\sigma_h^2}{MT^{2/5}}\Big),$$

where we define $F_t = \mathbb{E}[f(\boldsymbol{x}_t)] - f^\star$.

Choosing M such that it minimizes the right-hand side we get:

$$\min_{0 \leq t \leq T-1} F_t$$

$$= \mathcal{O}\Big(\frac{LD^3}{T^2} + \frac{L^{1/3}D^{7/3}\tilde{\sigma}_h^{2/3}}{T^{4/3}} + \frac{L^{1/5}D^{7/5}\sigma_g^{4/5}}{T^{2/5}}$$

$$+ F_0 e^{-T/3} + \frac{a_h^{3/5}D^2\sigma_h}{T^{6/5}} + \frac{a_g^{3/7}D\sigma_g}{T^{2/7}}\Big).$$

The last three terms in Theorem 3.5 correspond to the effect of the initialization and the initial gradient and Hessian noises. The last two terms can be made even smaller by choosing a larger initial batch (i.e. reducing a_g, a_h).

Again, we improve both Hessian and gradient rates and can converge for any precision $\varepsilon > 0$ for any batch size.

By ignoring the last three terms, we get the following iteration complexity: $\mathcal{O}\Big(\big(\frac{LD^3}{\varepsilon}\big)^{1/3} + \frac{L^{1/4}D^{7/4}\tilde{\sigma}_h^{1/2}}{\varepsilon^{3/4}} + \frac{L^{1/2}D^{7/2}\sigma_g^2}{\varepsilon^{5/2}}\Big).$

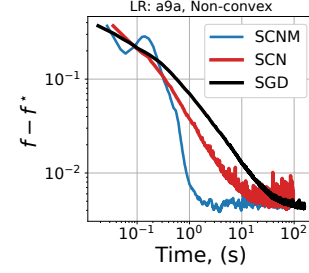where ψ is a simple possibly nonsmooth convex function.

Figure 1: A small batch size $b_g = 512, b_h = 216$ is used. We see SCNM, our Algorithm with momentum, outperforms its counterpart without momentum, it also outperforms SGD.
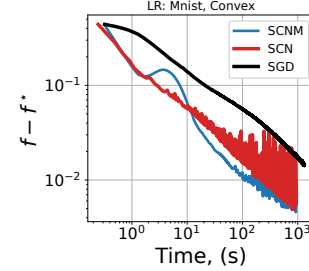

Figure 2: Logistic regression on the MNIST dataset. A small batch size $b_g = 512, b_h = 216$ is used. SCNM, our Algorithm with momentum, again outperforms its counterpart without momentum and SGD.

We note that in the convex case, it is possible to reach better rates for the Cubic Newton using acceleration Nesterov (2008). For example, (Agafonov et al., 2024) have a rate for their stochastic second-order method: $\mathcal{O}\Big(\big(\frac{LD^3}{\varepsilon}\big)^{1/3} + \frac{D\tilde{\sigma}_h^{1/2}}{\varepsilon^{1/2}} + \frac{D\sigma_g}{\varepsilon^2}\Big)$. We believe our momentum can be coupled with acceleration as well, and has the potential to improve this result further.

4 Experiments

We consider a logistic regression with a non-convex regularization ($Reg(\boldsymbol{x}) = \sum_{i=1}^d \frac{\boldsymbol{x}_i^2}{1+\boldsymbol{x}_i^2}$), following Kohler and Lucchi (2017). We use the A9A dataset from the LibSVM library (Chang and Lin, 2011). Results are shown in Figure 1. We also consider the MNIST dataset (LeCun and Cortes, 2010) and show the results in Figure 2. Both experiments depicted in Figures 1,2 use momentum parameter values $\alpha = 0.1$ and $\beta = 0.01$. We see that using momentum in the Stochastic Cubic Newton indeed improves its convergence rate and visibly decreases the variance, as predicted by our theory.

5 Discussion and Limitations

Practical implementation. We did not provide a specific method for solving the cubic sub-problem in

Algorithm 1. A naive implementation would require access to the entire Hessian matrix, making it impractical for high-dimensional problems. However, we can solve this problem using only Hessian-vector products, which are typically more efficient for various problem classes, such as neural networks with backpropagation. It is important to note that incorporating Hessian momentum may complicate this process. Therefore, we leave the exploration of how to effectively combine Hessian-vector products and momentum for future work.

Other momentum choices. As shown in Subsection 2.2, there are other types of momentum that we tried in the supplementary material. We were able to improve the gradient noise term from $\mathcal{O}(\frac{1}{T^{3/7}})$ in (16) to $\mathcal{O}(\frac{1}{T^{9/20}})$ using the MVR momentum from (Cutkosky and Orabona, 2020); this improvement comes at the cost of needing to assume the gradient of the objective is L_g-Lipschitz, but this constant can be made as big as we want by adding quadratic functions to our objective. Even when ignoring this problem with the constant L_g, the corresponding iteration complexity is only $\mathcal{O}(\frac{1}{\varepsilon^{30/9}})$, which is worse than $\mathcal{O}(\frac{1}{\varepsilon^3})$ obtained using the corresponding momentum in first-order methods (Cutkosky and Orabona, 2020). One of the reasons we think this is the case is that we did not have corresponding powers of r_t in the iterate bias that let us compensate for the bias using the term Mr_t^3 in Lemma 2.1. We note that this exact problem is present for first-order methods (although it is less complicated since we know r_t explicitly in this case) and the fix was to use normalized methods such as in (Cutkosky and Mehta, 2020b) to make the size of the step constant. Such a technique is still absent for second-order methods, and were it to be discovered, it would mean the possibility of improving the other types of momentum we discussed in this work.

Adaptive versions. The main drawback of our method is that it introduces an additional two hyperparameters α, β; while we can choose typical values as we did in the experiments, developing strategies to adapt to these new hyperparameters is important. We see our work as a first step in showcasing the potential benefits of using momentum in second-order methods and keep these questions for future research.

6 Conclusion

In this work, we study Stochastic Cubic Newton (SCN) with a specific type of momentum for non-convex problems and show improved rates compared to the plain version without momentum. Thus, we show, for the first time, the convergence of SCN for arbitrary batch sizes.

References

Agafonov, A., Kamzolov, D., Dvurechensky, P., Gasnikov, A., and Takáč, M. (2020). Inexact tensor methods and their application to stochastic convex optimization. *arXiv preprint arXiv:2012.15636*.

Agafonov, A., Kamzolov, D., Gasnikov, A., Kavis, A., Antonakopoulos, K., Cevher, V., and Takáč, M. (2024). Advancing the lower bounds: an accelerated, stochastic, second-order method with optimal adaptation to inexactness. In *The Twelfth International Conference on Learning Representations*.

Antonakopoulos, K., Kavis, A., and Cevher, V. (2022). Extra-Newton: A first approach to noise-adaptive accelerated second-order methods. *Advances in Neural Information Processing Systems*, 35:29859–29872.

Arnold, S. M. R., Manzagol, P.-A., Babanezhad, R., Mitliagkas, I., and Roux, N. L. (2019). Reducing the variance in online optimization by transporting past gradients.

Bennett, A. A. (1916). Newton's method in general analysis. *Proceedings of the National Academy of Sciences of the United States of America*, 2(10):592.

Benson, H. Y. and Shanno, D. F. (2018). Cubic regularization in symmetric rank-1 quasi-Newton methods. *Mathematical Programming Computation*, 10:457–486.

Cartis, C., Gould, N. I., and Toint, P. L. (2011a). Adaptive cubic regularisation methods for unconstrained optimization. Part I: motivation, convergence and numerical results. *Mathematical Programming*, 127(2):245–295.

Cartis, C., Gould, N. I., and Toint, P. L. (2011b). Adaptive cubic regularisation methods for unconstrained optimization. Part II: worst-case function- and derivative-evaluation complexity. *Mathematical programming*, 130(2):295–319.

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

Cartis, C. and Scheinberg, K. (2018). Global convergence rate analysis of unconstrained optimization methods based on probabilistic models. *Mathematical Programming*, 169:337–375.

Cauchy, A. (1847). Méthode générale pour la résolution des systemes d'équations simultanées. *Comp. Rend. Sci. Paris*, 25(1847):536–538.

Chang, C.-C. and Lin, C.-J. (2011). Libsvm: a library for support vector machines. *ACM transactions on intelligent systems and technology (TIST)*, 2(3):1–27.

Chayti, E. M., Jaggi, M., and Doikov, N. (2024). Unified convergence theory of stochastic and variance-reduced cubic newton methods. *Transactions on Machine Learning Research*.

Chen, R. Y., Gittens, A., and Tropp, J. A. (2012). The masked sample covariance estimator: An analysis via matrix concentration inequalities.

Conn, A. R., Gould, N. I., and Toint, P. L. (2000). *Trust region methods*. SIAM.

Cutkosky, A. and Mehta, H. (2020a). Momentum improves normalized sgd. In *International conference on machine learning*, pages 2260–2268. PMLR.

Cutkosky, A. and Mehta, H. (2020b). Momentum improves normalized sgd.

Cutkosky, A. and Orabona, F. (2020). Momentum-based variance reduction in non-convex sgd.

Doikov, N. and Grapiglia, G. N. (2023). First and zeroth-order implementations of the regularized Newton method with lazy approximated hessians. *arXiv preprint arXiv:2309.02412*.

Doikov, N. and Nesterov, Y. (2020). Convex optimization based on global lower second-order models. *Advances in Neural Information Processing Systems*, 33:16546–16556.

Doikov, N., Stich, S. U., and Jaggi, M. (2024). Spectral preconditioning for gradient methods on graded nonconvex functions. In *41th International Conference on Machine Learning (ICML 2024)*, number 41.

Fine, H. B. (1916). On Newton's method of approximation. *Proceedings of the National Academy of Sciences of the United States of America*, 2(9):546.

Gao, Y. and Ng, M. K. (2022). A momentum accelerated adaptive cubic regularization method for nonconvex optimization.

Ghadimi, S., Liu, H., and Zhang, T. (2017). Second-order methods with cubic regularization under inexact information. *arXiv preprint arXiv:1710.05782*.

Goujaud, B., Taylor, A., and Dieuleveut, A. (2023). Provable non-accelerations of the heavy-ball method. *arXiv preprint arXiv:2307.11291*.

Grapiglia, G. N., Gonçalves, M. L., and Silva, G. (2022). A cubic regularization of newton's method with finite difference hessian approximations. *Numerical Algorithms*, 90(2):607–630.

Griewank, A. (1981). The modification of Newton's method for unconstrained optimization by bounding cubic terms. Technical report, Technical report NA/12.

Hanzely, F., Doikov, N., Nesterov, Y., and Richtárik, P. (2020). Stochastic subspace cubic Newton method. In *International Conference on Machine Learning*, pages 4027–4038. PMLR.

Hillar, C. J. and Lim, L.-H. (2013). Most tensor problems are np-hard. *Journal of the ACM (JACM) 60 45*.

Kamzolov, D., Ziu, K., Agafonov, A., and Takác, M. (2023). Cubic regularized quasi-Newton methods. *arXiv preprint arXiv:2302.04987*.

Kantorovich, L. V. (1948). On Newton's method for functional equations. In *Dokl. Akad. Nauk SSSR*, volume 59, pages 1237–1240.

Kiefer, J. and Wolfowitz, J. (1952). Stochastic estimation of the maximum of a regression function. *Ann. Math. Statist. Volume 23, Number 3, 462-466*.

Kingma, D. P. and Ba, J. (2017). Adam: A method for stochastic optimization.

Kohler, J. M. and Lucchi, A. (2017). Sub-sampled cubic regularization for non-convex optimization. In *International Conference on Machine Learning*, pages 1895–1904. PMLR.

Korpelevich, G. M. (1976). The extragradient method for finding saddle points and other problems. *Matecon*, 12:747–756.

Lan, G. (2020). *First-order and stochastic optimization methods for machine learning*, volume 1. Springer.

LeCun, Y. and Cortes, C. (2010). MNIST handwritten digit database.

Marumo, N. and Takeda, A. (2024). Universal heavy-ball method for nonconvex optimization under hölder continuous hessians. *Mathematical Programming*, pages 1–29.

Nemirovski, A. (2004). Prox-method with rate of convergence o(1/t) for variational inequalities with lipschitz continuous monotone operators and smooth convex-concave saddle point problems. *SIAM Journal on Optimization*, 15(1):229–251.

Nesterov, Y. (1983). A method of solving a convex programming problem with convergence rate $O(1/k^2)$. 27(2):372–376.

Nesterov, Y. (2008). Accelerating the cubic regularization of Newton's method on convex problems. *Mathematical Programming*, 112(1):159–181.

Nesterov, Y. (2018). *Lectures on convex optimization*, volume 137. Springer.

Nesterov, Y. and Polyak, B. (2006). Cubic regularization of Newton's method and its global performance. *Mathematical Programming*, 108(1):177–205.

Nilesh, T., Mitchell, S., Chi, J., Jeffrey, R., and Michael I., J. (2018). Stochastic cubic regularization for fast nonconvex optimization. *Part of Advances in Neural Information Processing Systems 31*.

Nocedal, J. and Wright, S. (2006). *Numerical optimization*. Springer Science & Business Media.

Polyak, B. (1964). Some methods of speeding up the convergence of iteration methods. *USSR Computational Mathematics and Mathematical Physics*, 4(5):1–17.

Polyak, B. T. (1987). Introduction to optimization.

Polyak, B. T. (2007). Newton's method and its use in optimization. *European Journal of Operational Research*, 181(3):1086–1096.

Robbins, H. and Monro, S. (1951). A stochastic approximation method the annals of mathematical statistics. *Vol. 22, No. 3. pp. 400-407.*

Scieur, D. (2024). Adaptive quasi-Newton and anderson acceleration framework with explicit global (accelerated) convergence rates. In *International Conference on Artificial Intelligence and Statistics*, pages 883–891. PMLR.

Scieur, D. and Pedregosa, F. (2020). Universal average-case optimality of Polyak momentum. In *International conference on machine learning*, pages 8565–8572. PMLR.

Tran, H. and Cutkosky, A. (2022). Better SGD using second-order momentum. In Oh, A. H., Agarwal, A., Belgrave, D., and Cho, K., editors, *Advances in Neural Information Processing Systems*.

Tripuraneni, N., Stern, M., Jin, C., Regier, J., and Jordan, M. I. (2017). Stochastic cubic regularization for fast nonconvex optimization.

Wang, Z., Zhou, Y., Liang, Y., and Lan, G. (2018). A note on inexact condition for cubic regularized Newton's method. *arXiv preprint arXiv:1808.07384.*

Wang, Z., Zhou, Y., Liang, Y., and Lan, G. (2019). Cubic regularization with momentum for nonconvex optimization.

Xu, P., Roosta-Khorasani, F., and Mahoney, M. W. (2017). Newton-type methods for non-convex optimization under inexact Hessian information. *arXiv preprint arXiv:1708.07164 .*

Xu, P., Yang, J., Roosta-Khorasani, F., and Mahoney, M. W. (2016). Sub-sampled newton methods with non-uniform sampling.

Zhao, J., Lucchi, A., and Doikov, N. (2024). Cubic regularized subspace Newton for non-convex optimization. *arXiv preprint arXiv:2406.16666.*

Zhou, D., Xu, P., and Gu, Q. (2019). Stochastic variance-reduced cubic regularization methods. *Journal of Machine Learning Research 20 1-47.*

A Additional results using other momentums

We have also considered other combinations of momentum for the gradient estimation in the Cubic Newton method. We found out that they all guarantee convergence to arbitrary accuracy for any batch size. However, they make the Hessian term in the convergence rate slower, and only in the case of the MVR momentum we are able to show an improved rate for the gradient term.

We also note that to analyze the HB and MVR momentums, we needed to make additional assumptions on the smoothness of the objective function f. Specifically, for HB, we assumed that the gradients of f are L_g-Lipschitz, and for MVR, we needed to assume that the gradients of the sample functions (f_ξ) are all L_g-Lipschitz.

$$\text{(SOM) Section D:} \quad \boldsymbol{g}_t = (1 - \alpha)\Big(\boldsymbol{g}_{t-1} + \boldsymbol{H}_t(\boldsymbol{x}_t - \boldsymbol{x}_{t-1})\Big) + \alpha\nabla f_{\xi_t}(\boldsymbol{x}_t),$$

$$\text{(HB) Section E:} \quad \boldsymbol{g}_t = (1 - \alpha)\boldsymbol{g}_{t-1} + \alpha\nabla f_{\xi_t}(\boldsymbol{x}_t),$$

$$\text{(MVR) Section F:} \quad \boldsymbol{g}_t = (1 - \alpha)\Big(\boldsymbol{g}_{t-1} + \nabla_{\xi_t} f(\boldsymbol{x}_t) - \nabla_{\xi_t} f(\boldsymbol{x}_{t-1})\Big) + \alpha\nabla_{\xi_t} f(\boldsymbol{x}_t),$$

and for all these gradient estimators, we consider the HB momentum for the Hessian:

$$\boldsymbol{H}_t = (1 - \beta)\boldsymbol{H}_{t-1} + \beta\nabla^2 f_{\xi_t}(\boldsymbol{x}_t), \text{ for.}$$

To give a brief summary of the new results, let us assume for simplicity that $\sigma_{g,0} = 0$. Then we get the following rates after T iterations of our algorithm using these different types of momentum:

$$\mathbb{E}\mu_M(\boldsymbol{x}_{out}) = \begin{cases} \mathcal{O}\Big(\frac{\sqrt{L}F_0}{T} + \Big(\frac{F_0\sigma_h}{T}\Big)^{3/4} + \frac{\sigma_g^{9/16}\sigma_h^{3/8}F_0^{3/8}}{T^{3/8}} + \frac{L^{3/10}\sigma_g^{3/5}F_0^{3/5}}{T^{3/5}}\Big), & \text{(SOM) Section D} \\ \mathcal{O}\Big(\frac{\sqrt{L}F_0}{T} + \Big(\frac{F_0\tilde{\sigma}_h}{T}\Big)^{3/4} + \Big(\frac{L_g F_0\sigma_g^2}{T}\Big)^{3/8}\Big), & \text{(HB) Section E} \\ \mathcal{O}\Big(\frac{\sqrt{L}F_0}{T} + \Big(\frac{F_0\tilde{\sigma}_h}{T}\Big)^{3/4} + \Big(\frac{L_g F_0}{T}\Big)^{9/20}\sigma_g^{6/10}\Big), & \text{(MVR) Section F} \end{cases}$$

where $\tilde{\sigma}_h = \mathcal{O}(L_g + \sigma_h)$.

In particular, we see an improved dependence on T for the "gradient" term of the (MVR) momentum which is better than the $\frac{1}{T^{3/7}}$ that we got in the main paper with the (IT) momentum. However, this improved dependence comes at the cost of needing an additional assumption and a dependence on L_g instead of L (note that L_g can be made arbitrarily big by adding a quadratic to our objective function, while L will be unchanged).

B Preliminaries

We consider the general problem
$$\min_{\boldsymbol{x}\in\mathbb{R}^d} f(\boldsymbol{x}) := \mathbb{E}_\xi[f_\xi(\boldsymbol{x})],$$

where f_ξ is twice differentiable and f has L-Lipschitz Hessian i.e.:

$$\|\nabla^2 f(\boldsymbol{x}) - \nabla^2 f(\boldsymbol{y})\| \leq L\|\boldsymbol{x} - \boldsymbol{y}\|, \qquad \forall \boldsymbol{x}, \boldsymbol{y} \in \mathbb{R}^d. \tag{19}$$

As a direct consequence of (19) (see (Nesterov and Polyak, 2006; Nesterov, 2018)) we have for all $\boldsymbol{x}, \boldsymbol{y} \in \mathbb{R}^d$:

$$\|\nabla f(\boldsymbol{y}) - \nabla f(\boldsymbol{x}) - \nabla^2 f(\boldsymbol{x})(\boldsymbol{y} - \boldsymbol{x})\| \leq \frac{L}{2}\|\boldsymbol{x} - \boldsymbol{y}\|^2, \tag{20}$$

$$|f(\boldsymbol{y}) - f(\boldsymbol{x}) - \langle\nabla f(\boldsymbol{x}), \boldsymbol{y} - \boldsymbol{x}\rangle - \frac{1}{2}\langle\nabla^2 f(\boldsymbol{x})(\boldsymbol{y} - \boldsymbol{x}), \boldsymbol{y} - \boldsymbol{x}\rangle| \leq \frac{L}{6}\|\boldsymbol{y} - \boldsymbol{x}\|^3. \tag{21}$$

For a given vector $\boldsymbol{g}$ and matrix $\boldsymbol{H}$, we define $\boldsymbol{x}^+$ as

$$\boldsymbol{x}^+ \in \arg\min_{\boldsymbol{y}\in\mathbb{R}^d}\Big\{\Omega_{M,\boldsymbol{g},\boldsymbol{H}}(\boldsymbol{y}, \boldsymbol{x}) := \langle\boldsymbol{g}, \boldsymbol{y} - \boldsymbol{x}\rangle + \frac{1}{2}\langle\boldsymbol{H}(\boldsymbol{y} - \boldsymbol{x}), \boldsymbol{y} - \boldsymbol{x}\rangle + \frac{M}{6}\|\boldsymbol{y} - \boldsymbol{x}\|^3\Big\}. \tag{22}$$

We define the quantity

$$\mu_M(\boldsymbol{x}) = \max(\|\nabla f(\boldsymbol{x})\|^{3/2}, \frac{-\lambda_{min}(\nabla^2 f(\boldsymbol{x}))^3}{M^{3/2}}),$$

where $\lambda_{min}(\boldsymbol{A})$ denotes the smallest eigenvalue of the matrix $\boldsymbol{A}$.

Chayti et al. (2024) show the following theorem (see Appendix C, Theorem C.1):

Theorem B.1. *For any $\boldsymbol{x} \in \mathbb{R}^d$, let $\boldsymbol{x}^+$ be defined by (22). Then, for $M \geq L$ we have:*

$$f(\boldsymbol{x}) - f(\boldsymbol{x}^+) \geq \frac{1}{1008\sqrt{M}}\mu_M(\boldsymbol{x}^+) + \frac{M\|\boldsymbol{x}-\boldsymbol{x}^+\|^3}{72} - \frac{4\|\nabla f(\boldsymbol{x})-\boldsymbol{g}\|^{3/2}}{\sqrt{M}} - \frac{73\|\nabla^2 f(\boldsymbol{x})-\boldsymbol{H}\|^3}{M^2}.$$

We will also need the following Lemma to deal with averages under cubic powers:

Lemma B.2. *Suppose that $q \geq 2$, $p \geq 2$, and fix $r \geq \max(q, 2\log(p))$. Consider (weakly) independent random self-adjoint matrices $Y_1, \cdots, Y_N$ with dimension $p \times p$, $\mathbb{E}[Y_i] = 0$. It holds that:*

$$\left[\mathbb{E}[\|\sum_{i=1}^{N} Y_i\|^q]\right]^{1/q} \leq 2\sqrt{er}\|\left(\sum_{i=1}^{N}\mathbb{E}[Y_i^2]\right)^{1/2}\| + 4er\mathbb{E}[\max_i \|Y_i\|^q]^{1/q}.$$

The proof of this Lemma follows exactly the same lines as Lemma 32 from (Zhou et al., 2019) and Theorem A.1 from (Chen et al., 2012), and noticing that for the symmetrization trick to work, we only need the sequence to be (weakly) independent.

C Proofs of results of the main paper

C.1 Different forms of momentum

In the paper, we introduce the quantity

$$\Pi_t(\boldsymbol{x}) := \frac{1}{t}\sum_{j=0}^{t}\pi_{\xi_j}(\boldsymbol{x}) \ .$$

and define $\alpha_t = 1/t$.

Our aim is, at each step t, to approximate $\Pi_t(\boldsymbol{x}_t)$.

We also assume for illustration purposes that the functions π_ξ are all linear with the same gradient $\boldsymbol{A}$.

Then we have:

$$\begin{aligned}
\Pi_t(\boldsymbol{x}_t) &= (1-\alpha_t)\Pi_{t-1}(\boldsymbol{x}_t) + \alpha_t\pi_{\xi_t}(\boldsymbol{x}_t) \\
&= (1-\alpha_t)\Big[\Pi_{t-1}(\boldsymbol{x}_{t-1}) + \boldsymbol{A}(\boldsymbol{x}_t - \boldsymbol{x}_{t-1})\Big] + \alpha_t\pi_{\xi_t}(\boldsymbol{x}_t) \\
&= (1-\alpha_t)\Big[\Pi_{t-1}(\boldsymbol{x}_{t-1}) + \nabla\pi_{\xi_t}(\boldsymbol{x}_t)(\boldsymbol{x}_t - \boldsymbol{x}_{t-1})\Big] + \alpha_t\pi_{\xi_t}(\boldsymbol{x}_t) && \text{(SOM)} \\
&= (1-\alpha_t)\Big[\Pi_{t-1}(\boldsymbol{x}_{t-1}) + \pi_{\xi_t}(\boldsymbol{x}_t) - \pi_{\xi_t}(\boldsymbol{x}_{t-1})\Big] + \alpha_t\pi_{\xi_t}(\boldsymbol{x}_t) && \text{(MVR)} \\
&= (1-\alpha_t)\Pi_{t-1}(\boldsymbol{x}_{t-1}) + \alpha_t\Big(\pi_{\xi_t}(\boldsymbol{x}_t) + \frac{1-\alpha_t}{\alpha_t}\nabla\pi_{\xi_t}(\boldsymbol{x}_t)(\boldsymbol{x}_t - \boldsymbol{x}_{t-1})\Big) \\
&= (1-\alpha_t)\Pi_{t-1}(\boldsymbol{x}_{t-1}) + \alpha_t\pi_{\xi_t}\Big(\boldsymbol{x}_t + \frac{1-\alpha_t}{\alpha_t}(\boldsymbol{x}_t - \boldsymbol{x}_{t-1})\Big). && \text{(IT)}
\end{aligned}$$

C.2 Momentum bounds

The gradient and Hessian updates we used in the main paper have the following form:

$$\boldsymbol{g}_t = (1-\alpha)\boldsymbol{g}_{t-1} + \alpha\nabla f_{\xi_t}\Big(\boldsymbol{x}_t + \frac{1-\alpha}{\alpha}(\boldsymbol{x}_t - \boldsymbol{x}_{t-1})\Big), \text{ for } t \geq 1 \text{ and } \boldsymbol{g}_0 = \nabla f_{\xi_0}(\boldsymbol{x}_0) \tag{23}$$

and

$$\boldsymbol{H}_t = (1 - \beta)\boldsymbol{H}_{t-1} + \beta\nabla^2 f_{\xi_t}(\boldsymbol{x}_t), \text{ for } t \geq 1 \text{ and } \boldsymbol{H}_0 = \nabla^2 f_{\xi_0}(\boldsymbol{x}_0) \, , \tag{24}$$

for $\alpha, \beta \in (0, 1]$.

We define $r_t = \|\boldsymbol{x}_t - \boldsymbol{x}_{t-1}\|$.

Bounding the gradient momentum error. Let $\boldsymbol{\varepsilon}_t := \boldsymbol{g}_t - \nabla f(\boldsymbol{x}_t)$ and $\boldsymbol{z}(\boldsymbol{a}, \boldsymbol{b}) := \nabla f(\boldsymbol{a}) - \nabla f(\boldsymbol{b}) - \nabla^2 f(\boldsymbol{b})(\boldsymbol{a} - \boldsymbol{b})$.

Note that (20) implies $\|\boldsymbol{z}(\boldsymbol{a}, \boldsymbol{b})\| \leq \frac{L}{2}\|\boldsymbol{a} - \boldsymbol{b}\|^2$.

For ease of notation, we will also define $\boldsymbol{\theta}_t := \boldsymbol{x}_t + \frac{1-\alpha}{\alpha}(\boldsymbol{x}_t - \boldsymbol{x}_{t-1})$, $\boldsymbol{n}_t = \nabla f_{\xi_t}(\boldsymbol{\theta}_t) - \nabla f(\boldsymbol{\theta}_t)$ and $\boldsymbol{n}_0 = \nabla f_{\xi_0}(\boldsymbol{x}_0) - \nabla f(\boldsymbol{x}_0)$.

> **Lemma C.1.** *For $t \geq 1$:*
>
> $$\frac{1}{T}\sum_{t=0}^{T-1}\mathbb{E}\|\varepsilon_t\|^{3/2} \leq \frac{3}{\alpha T}\sigma_{g,0}^{3/2} + \sqrt{3}\alpha^{3/4}\sigma_g^{3/2} + \frac{\sqrt{3}(1-\alpha)^{3/2}L^{3/2}}{\alpha^3}\frac{1}{T}\sum_{t=0}^{T-1}\mathbb{E}r_t^3$$

Proof. For $t \geq 1$, we have:

$$
\begin{aligned}
\varepsilon_t &= (1 - \alpha)\boldsymbol{g}_{t-1} + \alpha\nabla f_{\xi_t}(\boldsymbol{\theta}_t) - \nabla f(\boldsymbol{x}_t) \\
&= (1-\alpha)\Big(\boldsymbol{g}_{t-1} - \nabla f(\boldsymbol{x}_t)\Big) + \alpha\Big(\nabla f_{\xi_t}(\boldsymbol{\theta}_t) - \nabla f(\boldsymbol{x}_t)\Big) \\
&= (1-\alpha)\Big(\varepsilon_{t-1} + \nabla f(\boldsymbol{x}_{t-1}) - \nabla f(\boldsymbol{x}_t)\Big) + \alpha\Big(\boldsymbol{n}_t + \nabla f(\boldsymbol{\theta}_t) - \nabla f(\boldsymbol{x}_t)\Big) \\
&= (1-\alpha)\Big(\varepsilon_{t-1} + \boldsymbol{z}(\boldsymbol{x}_{t-1}, \boldsymbol{x}_t)\Big) + \alpha\Big(\boldsymbol{n}_t + \nabla f(\boldsymbol{\theta}_t) - \nabla f(\boldsymbol{x}_t) + \frac{1-\alpha}{\alpha}\nabla^2 f(\boldsymbol{x}_t)(\boldsymbol{x}_{t-1} - \boldsymbol{x}_t)\Big) \\
&= (1-\alpha)\Big(\varepsilon_{t-1} + \boldsymbol{z}(\boldsymbol{x}_{t-1}, \boldsymbol{x}_t)\Big) + \alpha\Big(\boldsymbol{n}_t + \nabla f(\boldsymbol{\theta}_t) - \nabla f(\boldsymbol{x}_t) - \nabla^2 f(\boldsymbol{x}_t)(\boldsymbol{\theta}_t - \boldsymbol{x}_t)\Big) \\
&= (1-\alpha)\Big(\varepsilon_{t-1} + \boldsymbol{z}(\boldsymbol{x}_{t-1}, \boldsymbol{x}_t)\Big) + \alpha\Big(\boldsymbol{n}_t + \boldsymbol{z}(\boldsymbol{\theta}_t, \boldsymbol{x}_t)\Big) \\
&= (1-\alpha)\varepsilon_{t-1} + (1-\alpha)\boldsymbol{z}(\boldsymbol{x}_{t-1}, \boldsymbol{x}_t) + \alpha\boldsymbol{z}(\boldsymbol{\theta}_t, \boldsymbol{x}_t) + \alpha\boldsymbol{n}_t.
\end{aligned}
$$

Thus

$$\varepsilon_t = (1-\alpha)^t\varepsilon_0 + \sum_{i=0}^{t-1}(1-\alpha)^i\Big((1-\alpha)\boldsymbol{z}(\boldsymbol{x}_{t-i-1}, \boldsymbol{x}_{t-i}) + \alpha\boldsymbol{z}(\boldsymbol{\theta}_{t-i}, \boldsymbol{x}_{t-i})\Big) + \alpha\sum_{i=0}^{t-1}(1-\alpha)^i\boldsymbol{n}_{t-i} \, , \ t \geq 1,$$

and

$$
\begin{aligned}
\|\varepsilon_t\| &\leq (1-\alpha)^t\|\varepsilon_0\| + \sum_{i=0}^{t-1}(1-\alpha)^i\Big((1-\alpha)\|\boldsymbol{z}(\boldsymbol{x}_{t-i-1}, \boldsymbol{x}_{t-i})\| + \alpha\|\boldsymbol{z}(\boldsymbol{\theta}_{t-i}, \boldsymbol{x}_{t-i})\|\Big) + \alpha\|\sum_{i=0}^{t-1}(1-\alpha)^i\boldsymbol{n}_{t-i}\| \\
&\leq (1-\alpha)^t\|\varepsilon_0\| + \frac{L}{2}\sum_{i=0}^{t-1}(1-\alpha)^i\Big((1-\alpha)\|\boldsymbol{x}_{t-i-1} - \boldsymbol{x}_{t-i}\|^2 + \alpha\|\boldsymbol{\theta}_{t-i} - \boldsymbol{x}_{t-i}\|^2\Big) + \alpha\|\sum_{i=0}^{t-1}(1-\alpha)^i\boldsymbol{n}_{t-i}\| \\
&\leq (1-\alpha)^t\|\varepsilon_0\| + \frac{L}{2}\sum_{i=0}^{t-1}(1-\alpha)^i\Big((1-\alpha)r_{t-i}^2 + \frac{(1-\alpha)^2}{\alpha}r_{t-i}^2\Big) + \alpha\|\sum_{i=0}^{t-1}(1-\alpha)^i\boldsymbol{n}_{t-i}\| \\
&\leq (1-\alpha)^t\|\varepsilon_0\| + \frac{(1-\alpha)L}{\alpha}\sum_{i=0}^{t-1}(1-\alpha)^ir_{t-i}^2 + \alpha\|\sum_{i=0}^{t-1}(1-\alpha)^i\boldsymbol{n}_{t-i}\|.
\end{aligned}
$$

Now we raise this to the power $3/2$ and using the convexity of $x \mapsto x^{3/2}$, we get:

$$\|\varepsilon_t\|^{3/2} \leq \sqrt{3}\Bigg((1-\alpha)^{3t/2}\|\varepsilon_0\|^{3/2} + \frac{(1-\alpha)^{3/2}L^{3/2}}{\alpha^{3/2}}\Big(\sum_{i=0}^{t-1}(1-\alpha)^ir_{t-i}^2\Big)^{3/2} + \alpha^{3/2}\|\sum_{i=0}^{t-1}(1-\alpha)^i\boldsymbol{n}_{t-i}\|^{3/2}\Bigg).$$

Again by the convexity of $x \mapsto x^{3/2}$, we have:

$$\Big(\sum_{i=0}^{t-1}(1-\alpha)^i r_{t-i}^2\Big)^{3/2} \leq \Big(\sum_{i=0}^{t-1}(1-\alpha)^i\Big)^{1/2}\sum_{i=0}^{t-1}(1-\alpha)^i r_{t-i}^3$$

$$\leq (\frac{1}{\alpha})^{1/2}\sum_{i=0}^{t-1}(1-\alpha)^i r_{t-i}^3.$$

Thus

$$\|\varepsilon_t\|^{3/2} \leq \sqrt{3}\bigg((1-\alpha)^{3t/2}\|\varepsilon_0\|^{3/2} + \frac{(1-\alpha)^{3/2}L^{3/2}}{\alpha^2}\sum_{i=0}^{t-1}(1-\alpha)^i r_{t-i}^3 + \alpha^{3/2}\|\sum_{i=0}^{t-1}(1-\alpha)^i \boldsymbol{n}_{t-i}\|^{3/2}\bigg).$$

Now we take the expectation over $(\xi_j)_{j \leq t}$, we get:

$$\mathbb{E}\|\varepsilon_t\|^{3/2} \leq \sqrt{3}\bigg((1-\alpha)^{3t/2}\mathbb{E}\|\boldsymbol{n}_0\|^{3/2} + \frac{(1-\alpha)^{3/2}L^{3/2}}{\alpha^2}\sum_{i=0}^{t-1}(1-\alpha)^i\mathbb{E}r_{t-i}^3 + \alpha^{3/2}\mathbb{E}\|\sum_{i=0}^{t-1}(1-\alpha)^i\boldsymbol{n}_{t-i}\|^{3/2}\bigg).$$

Assuming $\mathbb{E}\nabla f_\xi(\boldsymbol{x}_j) = \nabla f_\xi(\boldsymbol{x}_j)$ and $\mathbb{E}[\|\nabla f_\xi(\boldsymbol{x}_j) - \nabla f_\xi(\boldsymbol{x}_j)\|^2|\boldsymbol{x}_j] \leq \sigma_g^2$ and that ξ_j is drawn independently from the past, then, by Jensen's inequality, we have: $\mathbb{E}\|\boldsymbol{n}_0\|^{3/2} \leq \sigma_{g,0}^{3/2}$ and

$$\mathbb{E}\|\sum_{i=0}^{t-1}(1-\alpha)^i\boldsymbol{n}_{t-i}\|^{3/2} \leq \Big(\mathbb{E}\|\sum_{i=0}^{t-1}(1-\alpha)^i\boldsymbol{n}_{t-i}\|^2\Big)^{3/4}$$

$$= \Big(\sum_{i=0}^{t-1}(1-\alpha)^{2i}\mathbb{E}\|\boldsymbol{n}_{t-i}\|^2\Big)^{3/4}$$

$$\leq \Big(\sum_{i=0}^{t-1}(1-\alpha)^{2i}\sigma_g^2\Big)^{3/4}$$

$$\leq \Big(\sum_{i=0}^{t-1}(1-\alpha)^i\Big)^{3/4}\sigma_g^{3/2}$$

$$\leq \Big(\frac{1}{\alpha}\Big)^{3/4}\sigma_g^{3/2}.$$

Which means:

$$\mathbb{E}\|\varepsilon_t\|^{3/2} \leq \sqrt{3}\Big((1-\alpha)^{3t/2}\sigma_{g,0}^{3/2} + \frac{(1-\alpha)^{3/2}L^{3/2}}{\alpha^2}\sum_{i=0}^{t-1}(1-\alpha)^i\mathbb{E}r_{t-i}^3 + \alpha^{3/4}\sigma_g^{3/2}\Big)$$

Finally,

$$\frac{1}{T}\sum_{t=0}^{T-1}\mathbb{E}\|\varepsilon_t\|^{3/2} \leq \frac{\mathbb{E}\|\varepsilon_0\|^{3/2}}{T} + \frac{\sqrt{3}}{T}\sum_{t=1}^{T-1}\Big((1-\alpha)^{3t/2}\sigma_{g,0}^{3/2} + \frac{(1-\alpha)^{3/2}L^{3/2}}{\alpha^2}\sum_{i=0}^{t-1}(1-\alpha)^i\mathbb{E}r_{t-i}^3 + \alpha^{3/4}\sigma_g^{3/2}\Big)$$

$$\leq \frac{1}{T}\sigma_{g,0}^{3/2} + \frac{\sqrt{3}}{\alpha T}\sigma_{g,0}^{3/2} + \sqrt{3}\alpha^{3/4}\sigma_g^{3/2} + \frac{\sqrt{3}}{T}\frac{(1-\alpha)^{3/2}L^{3/2}}{\alpha^2}\sum_{t=0}^{T-1}\sum_{i=0}^{t-1}(1-\alpha)^i\mathbb{E}r_{t-i}^3$$

$$\leq \frac{3}{\alpha T}\sigma_{g,0}^{3/2} + \sqrt{3}\alpha^{3/4}\sigma_g^{3/2} + \frac{\sqrt{3}(1-\alpha)^{3/2}L^{3/2}}{\alpha^3}\frac{1}{T}\sum_{t=0}^{T-1}\mathbb{E}r_t^3$$

Which finishes the proof. $\qquad\square$

Bounding the Hessian momentum error.

We will do the same for the sequence $(\boldsymbol{H}_t)$. Let's define $\boldsymbol{\Sigma}_t = \boldsymbol{H}_t - \nabla^2 f(\boldsymbol{x}_t)$ and $\boldsymbol{Z}(\boldsymbol{a}, \boldsymbol{b}) = \nabla^2 f(\boldsymbol{a}) - \nabla^2 f(\boldsymbol{b})$ and note that by (19), we have $\|\boldsymbol{Z}(\boldsymbol{a}, \boldsymbol{b})\| \leq L\|\boldsymbol{a} - \boldsymbol{b}\|$.

Lemma C.2. *We show :*

$$\frac{1}{T}\sum_{t=0}^{T-1} \mathbb{E}\|\boldsymbol{\Sigma}_t\|^3 \leq \frac{10}{\beta T}\sigma_{h,0}^3 + \beta^{3/2}\tilde{\sigma}_h^3(\beta) + \frac{9(1-\beta)L^3}{\beta^3}\frac{1}{T}\sum_{t=0}^{T-1} \mathbb{E}r_t^3$$

where $\tilde{\sigma}_h^3(\beta) = 8(er)^{3/2}\sigma_h^3 + 16(er)^3\beta^{3/2}\delta_h^3$ *and* $r = 2\log(d)$.

Proof. We have

$$
\begin{aligned}
\boldsymbol{\Sigma}_t &= (1-\beta)\boldsymbol{H}_{t-1} + \beta\nabla^2 f_{\xi_t}(\boldsymbol{x}_t) - \nabla^2 f(\boldsymbol{x}_t) \\
&= (1-\beta)\Big(\boldsymbol{H}_{t-1} - \nabla^2 f(\boldsymbol{x}_t)\Big) + \beta\Big(\nabla^2 f_{\xi_t}(\boldsymbol{x}_t) - \nabla^2 f(\boldsymbol{x}_t)\Big) \\
&= (1-\beta)\Big(\boldsymbol{\Sigma}_{t-1} + \nabla^2 f(\boldsymbol{x}_{t-1}) - \nabla^2 f(\boldsymbol{x}_t)\Big) + \beta\Big(\nabla^2 f_{\xi_t}(\boldsymbol{x}_t) - \nabla^2 f(\boldsymbol{x}_t)\Big) \\
&= (1-\beta)\boldsymbol{\Sigma}_{t-1} + (1-\beta)\boldsymbol{Z}(\boldsymbol{x}_{t-1}, \boldsymbol{x}_t) + \beta\boldsymbol{N}_t
\end{aligned}
$$

Where we defined the $\boldsymbol{N}_t := \nabla^2 f_{\xi_t}(\boldsymbol{x}_t) - \nabla^2 f(\boldsymbol{x}_t)$.

Thus

$$\boldsymbol{\Sigma}_t = (1-\beta)^t\boldsymbol{\Sigma}_0 + (1-\beta)\sum_{i=0}^{t-1}\boldsymbol{Z}(\boldsymbol{x}_{t-i-1}, \boldsymbol{x}_{t-i}) + \beta\sum_{i=0}^{t-1}\boldsymbol{N}_{t-i}$$

we take the norm now:

$$
\begin{aligned}
\|\boldsymbol{\Sigma}_t\| &\leq (1-\beta)^t\|\boldsymbol{\Sigma}_0\| + (1-\beta)\sum_{i=0}^{t-1}(1-\beta)^i\|\boldsymbol{Z}(\boldsymbol{x}_{t-i-1}, \boldsymbol{x}_{t-i})\| + \beta\|\sum_{i=0}^{t-1}(1-\beta)^i\boldsymbol{N}_{t-i}\| \\
&\leq (1-\beta)^t\|\boldsymbol{\Sigma}_0\| + (1-\beta)L\sum_{i=0}^{t-1}(1-\beta)^i r_{t-i} + \beta\|\sum_{i=0}^{t-1}(1-\beta)^i\boldsymbol{N}_{t-i}\|
\end{aligned}
$$

We use the convexity of $x \mapsto x^3$ and get :

$$
\begin{aligned}
\|\boldsymbol{\Sigma}_t\|^3 &\leq 9\Big((1-\beta)^{3t}\|\boldsymbol{\Sigma}_0\|^3 + (1-\beta)L^3(\sum_{i=0}^{t-1}(1-\beta)^i r_{t-i})^3 + \beta^3\|\sum_{i=0}^{t-1}(1-\beta)^i\boldsymbol{N}_{t-i}\|^3\Big) \\
&\leq 9\Big((1-\beta)^{3t}\|\boldsymbol{N}_0\|^3 + \frac{(1-\beta)L^3}{\beta^2}\sum_{i=0}^{t-1}(1-\beta)^i r_{t-i}^3 + \beta^3\|\sum_{i=0}^{t-1}(1-\beta)^i\boldsymbol{N}_{t-i}\|^3\Big)
\end{aligned}
$$

We take the expectation and get

$$\mathbb{E}\|\boldsymbol{\Sigma}_t\|^3 \leq 9\Big((1-\beta)^{3t}\mathbb{E}\|\boldsymbol{N}_0\|^3 + \frac{(1-\beta)L^3}{\beta^2}\sum_{i=0}^{t-1}(1-\beta)^i\mathbb{E}r_{t-i}^3 + \beta^3\mathbb{E}\|\sum_{i=0}^{t-1}(1-\beta)^i\boldsymbol{N}_{t-i}\|^3\Big) \qquad (25)$$

We make the following assumptions about the Hessian noises $\boldsymbol{N}_t$:

- the noise is centered: $\mathbb{E}\boldsymbol{N}_t = 0$;
- the noise has a bounded variance: there exists $\sigma_h \geq 0$ such that $\mathbb{E}\|\boldsymbol{N}_t\|^2 \leq \sigma_h$;
- the noise is bounded almost surely, i.e. there exists $\delta_h \geq \sigma_h$ such that $\|\boldsymbol{N}_t\| \leq \delta_h$ with probability one.

We also assume that the noise at the start $\mathbb{E}\|\boldsymbol{N}_0\|^2 \leq \sigma_{h,0}^2$ where $\sigma_{h,0}$ is potentially much smaller than σ_h (e.g., by using a large batch at the start).

Note that all these noises are independent from each other conditioned on the past.

Under these assumptions we can show that: $\mathbb{E}\|\boldsymbol{N}_0\|^3 \leq \sigma_{h,0}^3$ and using the following Lemma B.2, we have for $q = 3$ and $r = 2log(d)$ where d is the dimension of the problem:

$$
\begin{aligned}
\Big(\mathbb{E}\|\sum_{i=0}^{t-1}(1-\beta)^i \boldsymbol{N}_{t-i}\|^3\Big)^{1/3} &\leq 2\sqrt{er}\|\big(\sum_{i=0}^{t-1}(1-\beta)^i\mathbb{E}[\boldsymbol{N}_{t-i}^2]\big)^{1/2}\| + 4er\mathbb{E}[\max_i\|(1-\beta)^i\boldsymbol{N}_{t-i}\|^q]^{1/q} \\
&\leq 2\sqrt{er}\|\big(\sum_{i=0}^{t-1}(1-\beta)^i\mathbb{E}[\boldsymbol{N}_{t-i}^2]\big)\|^{1/2} + 4er\mathbb{E}[\max_i\|(1-\beta)^i\boldsymbol{N}_{t-i}\|^q]^{1/q} \\
&\leq 2\sqrt{er}\big(\sum_{i=0}^{t-1}(1-\beta)^i\big)^{1/2}\sigma_h + 4er\delta_h \\
&\leq 2\sqrt{\frac{er}{\beta}}\sigma_h + 4er\delta_h.
\end{aligned}
$$

We conclude that :

$$
\begin{aligned}
\beta^3\mathbb{E}\|\sum_{i=0}^{t-1}(1-\beta)^i \boldsymbol{N}_{t-i}\|^3\beta^3 &\leq \Big(2\sqrt{\frac{er}{\beta}}\sigma_h + 4er\delta_h\Big)^3 \\
&\leq 8(er\beta)^{3/2}\sigma_h^3 + 16(er\beta)^3\delta_h^3 \\
&= \beta^{3/2}\Big(8(er)^{3/2}\sigma_h^3 + 16(er)^3\beta^{3/2}\delta_h^3\Big) \\
&:= \beta^{3/2}\tilde{\sigma}_h^3(\beta).
\end{aligned}
$$

Replacing this in (25), we get:

$$
\mathbb{E}\|\boldsymbol{\Sigma}_t\|^3 \leq 9\Big((1-\beta)^{3t}\sigma_{h,0}^3 + \beta^{3/2}\tilde{\sigma}_h^3(\beta) + \frac{(1-\beta)L^3}{\beta^2}\sum_{i=0}^{t-1}(1-\beta)^i\mathbb{E}r_{t-i}^3\Big).
$$

Finally:

$$
\begin{aligned}
\frac{1}{T}\sum_{t=0}^{T-1}\mathbb{E}\|\boldsymbol{\Sigma}_t\|^3 &\leq \frac{\mathbb{E}\|\boldsymbol{\Sigma}_0\|^3}{T} + \frac{9}{T}\sum_{t=1}^{T-1}\Big((1-\beta)^{3t}\sigma_{h,0}^3 + \beta^{3/2}\tilde{\sigma}_h^3(\beta) + \frac{(1-\beta)L^3}{\beta^2}\sum_{i=0}^{t-1}(1-\beta)^i\mathbb{E}r_{t-i}^3\Big) \\
&\leq \frac{1}{T}\sigma_{h,0}^3 + \frac{9}{\beta T}\sigma_{h,0}^3 + \beta^{3/2}\tilde{\sigma}_h^3(\beta) + \frac{9}{T}\frac{(1-\beta)L^3}{\beta^2}\sum_{t=0}^{T-1}\sum_{i=0}^{t-1}(1-\beta)^i\mathbb{E}r_{t-i}^3 \\
&\leq \frac{10}{\beta T}\sigma_{h,0}^3 + \beta^{3/2}\tilde{\sigma}_h^3(\beta) + \frac{9(1-\beta)L^3}{\beta^3}\frac{1}{T}\sum_{t=0}^{T-1}\mathbb{E}r_t^3.
\end{aligned}
$$

□

C.3 Nonconvex convergence

Using Theorem B.1, we have:

$$
\begin{aligned}
f(\boldsymbol{x}_t) - f(\boldsymbol{x}_{t+1}) &\geq \frac{1}{1008\sqrt{M}}\mu_M(\boldsymbol{x}_{t+1}) + \frac{M\|\boldsymbol{x}_t - \boldsymbol{x}_{t+1}\|^3}{72} - \frac{4\|\nabla f(\boldsymbol{x}_t) - \boldsymbol{g}_t\|^{3/2}}{\sqrt{M}} - \frac{73\|\nabla^2 f(\boldsymbol{x}_t) - \boldsymbol{H}_t\|^3}{M^2} \\
&\geq \frac{1}{1008\sqrt{M}}\mu_M(\boldsymbol{x}_{t+1}) + \frac{Mr_t^3}{72} - \frac{4\|\boldsymbol{\varepsilon}_t\|^{3/2}}{\sqrt{M}} - \frac{73\|\boldsymbol{\Sigma}_t\|^3}{M^2}.
\end{aligned}
$$

Reorganizing the inequality and taking the expectation we get:

$$\frac{1}{1008}\mathbb{E}\mu_M(\boldsymbol{x}_{t+1}) \leq \sqrt{M}\mathbb{E}(f(\boldsymbol{x}_t) - f(\boldsymbol{x}_{t+1})) - \frac{M^{3/2}\mathbb{E}r_t^3}{72} + 4\mathbb{E}\|\boldsymbol{\varepsilon}_t\|^{3/2} + \frac{73\mathbb{E}\|\boldsymbol{\Sigma}_t\|^3}{M^{3/2}}.$$

Summing over $t = 0, T-1$ and dividing by T, we get:

$$\frac{1}{1008}\frac{1}{T}\sum_{t=0}^{T-1}\mathbb{E}\mu_M(\boldsymbol{x}_{t+1}) \leq \frac{\sqrt{M}F_0}{T} - \frac{M^{3/2}}{72}\mathbb{E}\frac{1}{T}\sum_{t=0}^{T-1}r_t^3 + 4\frac{1}{T}\sum_{t=0}^{T-1}\mathbb{E}\|\boldsymbol{\varepsilon}_t\|^{3/2} + \frac{73}{M^{3/2}}\frac{1}{T}\sum_{t=0}^{T-1}\mathbb{E}\|\boldsymbol{\Sigma}_t\|^3,$$

where $F_0 := f(\boldsymbol{x}_0) - \min_{\boldsymbol{x}} f(\boldsymbol{x})$ and we assume $F_0 < +\infty$.

We have already shown that :

$$\frac{1}{T}\sum_{t=0}^{T-1}\mathbb{E}\|\boldsymbol{\varepsilon}_t\|^{3/2} \leq \frac{3}{\alpha T}\sigma_{g,0}^{3/2} + \sqrt{3}\alpha^{3/4}\sigma_g^{3/2} + \frac{\sqrt{3}(1-\alpha)^{3/2}L^{3/2}}{\alpha^3}\frac{1}{T}\sum_{t=0}^{T-1}\mathbb{E}r_t^3$$

and

$$\frac{1}{T}\sum_{t=0}^{T-1}\mathbb{E}\|\boldsymbol{\Sigma}_t\|^3 \leq \frac{10}{\beta T}\sigma_{h,0}^3 + \beta^{3/2}\tilde{\sigma}_h^3(\beta) + \frac{9(1-\beta)L^3}{\beta^3}\frac{1}{T}\sum_{t=0}^{T-1}\mathbb{E}r_t^3.$$

Thus we get:

$$\frac{1}{1008}\frac{1}{T}\sum_{t=0}^{T-1}\mathbb{E}\mu_M(\boldsymbol{x}_{t+1}) \leq \frac{\sqrt{M}F_0}{T} - \frac{M^{3/2}}{72}\mathbb{E}\frac{1}{T}\sum_{t=0}^{T-1}r_t^3 + 4\frac{1}{T}\sum_{t=0}^{T-1}\mathbb{E}\|\boldsymbol{\varepsilon}_t\|^{3/2} + \frac{73}{M^{3/2}}\frac{1}{T}\sum_{t=0}^{T-1}\mathbb{E}\|\boldsymbol{\Sigma}_t\|^3$$

$$\leq \frac{\sqrt{M}F_0}{T} + \frac{12}{\alpha T}\sigma_{g,0}^{3/2} + 4\sqrt{3}\alpha^{3/4}\sigma_g^{3/2} + \frac{730}{M^{3/2}\beta T}\sigma_{h,0}^3 + \frac{73}{M^{3/2}}\beta^{3/2}\tilde{\sigma}_h^3(\beta)$$

$$+ \left(\frac{4\sqrt{3}L^{3/2}}{\alpha^3} + \frac{73}{M^{3/2}}\frac{9L^3}{\beta^3} - \frac{M^{3/2}}{72}\right)\mathbb{E}\frac{1}{T}\sum_{t=0}^{T-1}r_t^3.$$

We choose $\alpha, \beta \in (0,1]$ such that:

$$\frac{4\sqrt{3}L^{3/2}}{\alpha^3} + \frac{73}{M^{3/2}}\frac{9L^3}{\beta^3} - \frac{M^{3/2}}{72} \leq 0 , \tag{26}$$

and to deal with initial gradient and Hessian noises we also take:

$\alpha \in \arg\min \frac{12}{\alpha T}\sigma_{g,0}^{3/2} + 4\sqrt{3}\alpha^{3/4}\sigma_g^{3/2} \approx \frac{a_g^{6/7}}{T^{4/7}}$ and $\beta \in \arg\min \frac{730}{M^{3/2}\beta T}\sigma_{h,0}^3 + \frac{73}{M^{3/2}}\beta^{3/2}\sigma_h^3 \approx \frac{a_h^{6/5}}{T^{2/5}}$.

Where, for simplicity, we denoted $a_g = \frac{\sigma_{g,0}}{\sigma_g} \leq 1$ and $a_h = \frac{\sigma_{h,0}}{\sigma_h} \leq 1$.

All in all, we choose

$$\alpha = \max\left(\frac{a_g^{6/7}}{T^{4/7}}, 10\sqrt{\frac{L}{M}}\right), \ \beta = \max\left(\frac{a_h^{6/5}}{T^{2/5}}, 46\frac{L}{M}\right). \tag{27}$$

To ensure $\alpha, \beta \leq 1$ it suffices to take $M \geq 100L$.

Finally, we get:

$$\frac{1}{1008}\frac{1}{T}\sum_{t=0}^{T-1}\mathbb{E}\mu_M(\boldsymbol{x}_{t+1}) \leq \frac{\sqrt{M}F_0}{T} + 40\frac{L^{3/8}}{M^{3/8}}\sigma_g^{3/2} + \frac{73(46)^{3/2}L^{3/2}}{M^3}\tilde{\sigma}_h^3(\beta) + 20\frac{a_g^{9/14}\sigma_g^{3/2}}{T^{3/7}} + 803\frac{a_h^{9/5}\sigma_h^3}{M^{3/2}T^{3/5}},$$

where $\tilde{\sigma}_h^3(\beta) = 8(er)^{3/2}\sigma_h^3 + 16(46)^{3/2}(er)^3\frac{L^{3/2}}{M^{3/2}}\delta_h^3$ (remember that $r = 2\log(d)$).

Compare this to the bound we get without using momentum:

$$\mathcal{O}\Big(\frac{\sqrt{M}F_0}{T} + \sigma_g^{3/2} + \frac{1}{M^{3/2}}\sigma_h^3\Big).$$

We clearly depend better on the regularization parameter M in both the gradient and Hessian noise terms.

Going back to the bound

$$\frac{1}{1008}\frac{1}{T}\sum_{t=0}^{T-1}\mathbb{E}\mu_M(\boldsymbol{x}_{t+1}) \leq \frac{\sqrt{M}F_0}{T} + 40\frac{L^{3/8}}{M^{3/8}}\sigma_g^{3/2} + \frac{73(46)^{3/2}L^{3/2}}{M^3}\tilde{\sigma}_h^3(\beta) + 20\frac{a_g^{9/14}\sigma_g^{3/2}}{T^{3/7}} + 803\frac{a_h^{9/5}\sigma_h^3}{M^{3/2}T^{3/5}}.$$

By choosing

$$M = \max\Big(100L, \frac{L^{3/7}\sigma_g^{12/7}}{F_0^{8/7}}T^{8/7}, \frac{L^{3/7}\tilde{\sigma}_h(T)^{3/7}}{F_0^{2/7}}T^{2/7}, \frac{a_h^{9/10}\sigma_h^{3/2}T^{1/5}}{\sqrt{F_0}}\Big)$$

for $\tilde{\sigma}_h^3(T) = 8(er)^{3/2}\sigma_h^3 + 16(46)^{3/2}(er)^3\frac{L^{6/7}\tilde{F}_0^{3/7}}{\sigma_h^{9/7}T^{3/7}}\delta_h^3$, we get:

$$\frac{1}{1008}\frac{1}{T}\sum_{t=0}^{T-1}\mathbb{E}\mu_M(\boldsymbol{x}_{t+1}) \leq \frac{\sqrt{L}F_0}{T} + \frac{40L^{3/14}F_0^{3/7}\sigma_g^{6/7} + 20a_g^{9/14}\sigma_g^{3/2}}{T^{3/7}} + 73(46)^{3/2}\frac{L^{3/14}F_0^{6/7}\tilde{\sigma}_h(T)^{3/7}}{T^{6/7}} + 804\frac{a_h^{9/20}\sigma_h^{3/4}F_0^{1/2}}{T^{9/10}}.$$

C.4 Convex case

We consider here the slightly more general stochastic composite optimization problem

$$\min_{\boldsymbol{x}} f(\boldsymbol{x}) := \underbrace{\mathbb{E}_{\xi\sim\mathcal{P}}\Big[f_\xi(\boldsymbol{x})\Big]}_{:=\tilde{f}(\boldsymbol{x})} + \psi(\boldsymbol{x}) , \tag{28}$$

where ψ is a simple, convex, potentially non-smooth function and f is convex. Simple here means that we can efficiently solve the following general subproblem:

$$\boldsymbol{x}_{t+1} \in \arg\min_{\boldsymbol{y}\in\mathbb{R}^d}\Big\{\Omega_{M,\boldsymbol{g}_t,\boldsymbol{H}_t}(\boldsymbol{y},\boldsymbol{x}_t) + \psi(\boldsymbol{y})\Big\}. \tag{29}$$

For example, ψ can be the indicator function of a bounded convex set, in this case, solving (29) has the same cost as solving the cubic subproblem (22).

The optimality condition of (29) guarantees that there exists a subgradient of ψ that we denote as $\psi'(\boldsymbol{x}_{t+1}) \in \partial\psi(\boldsymbol{x}_{t+1})$ such that

$$\boldsymbol{g}_t + \boldsymbol{H}_t(\boldsymbol{x}_{t+1} - \boldsymbol{x}_t) + \frac{M}{2}r_t^2 + \psi'(\boldsymbol{x}_{t+1}) = \boldsymbol{0}.$$

We define $\nabla f(\boldsymbol{x}_{t+1}) = \nabla\tilde{f}(\boldsymbol{x}_{t+1}) + \psi'(\boldsymbol{x}_{t+1})$.

We assume that the diameter of the domain of f is bounded by D. This means that :

$$\forall\boldsymbol{x}\in dom(f) \quad : \quad f(\boldsymbol{x}) - f^\star \leq D\|\nabla f(\boldsymbol{x})\| .$$

Following the same lines as Theorem C.1 from (Chayti et al., 2024), it is easy to show the same inequality as the one we had before for the progress:

$$f(\boldsymbol{x}_t) - f(\boldsymbol{x}_{t+1}) \geq \frac{1}{1008\sqrt{M}}\|\nabla f(\boldsymbol{x}_{t+1})\|^{3/2} + \frac{Mr_t^3}{72} - \frac{4\|\boldsymbol{\varepsilon}_t\|^{3/2}}{\sqrt{M}} - \frac{73\|\boldsymbol{\Sigma}_t\|^3}{M^2}$$

$$\geq \frac{1}{1008\sqrt{M}D^{3/2}}(f(\boldsymbol{x}_{t+1} - f^\star)^{3/2} + \frac{Mr_t^3}{72} - \frac{4\|\boldsymbol{\varepsilon}_t\|^{3/2}}{\sqrt{M}} - \frac{73\|\boldsymbol{\Sigma}_t\|^3}{M^2}.$$

Let us define $F_t = \mathbb{E}f(\boldsymbol{x}_t) - f^\star$, using Jensen's inequality we get

$$F_t - F_{t+1} \geq \frac{1}{1008\sqrt{M}D^{3/2}}F_{t+1}^{3/2} + \underbrace{\mathbb{E}\left[\frac{Mr_t^3}{72} - \frac{4\|\boldsymbol{\varepsilon}_t\|^{3/2}}{\sqrt{M}} - \frac{73\|\boldsymbol{\Sigma}_t\|^3}{M^2}\right]}_{:=-a_t}. \tag{30}$$

From the non-convex case, we have that

$$\frac{\sqrt{M}}{T}\sum_{t=0}^{T-1}a_t \leq \frac{12}{\alpha T}\sigma_{g,0}^{3/2} + 4\sqrt{3}\alpha^{3/4}\sigma_g^{3/2} + \frac{730}{M^{3/2}\beta T}\sigma_{h,0}^3 + \frac{73}{M^{3/2}}\beta^{3/2}\tilde{\sigma}_h^3(\beta)$$

under the condition that

$$\frac{4\sqrt{3}L^{3/2}}{\alpha^3} + \frac{73}{M^{3/2}}\frac{9L^3}{\beta^3} - \frac{M^{3/2}}{72} \leq 0 .$$

Thus using the same choices of α, β as in the non-convex case, we get:

$$\frac{\sqrt{M}}{T}\sum_{t=0}^{T-1}a_t \leq 20\frac{a_g^{9/14}\sigma_g^{3/2}}{T^{3/7}} + 803\frac{a_h^{9/5}\sigma_h^3}{M^{3/2}T^{3/5}} + 40\frac{L^{3/8}}{M^{3/8}}\sigma_g^{3/2} + \frac{73(46)^{3/2}L^{3/2}}{M^3}\tilde{\sigma}_h^3(\beta) , \tag{31}$$

Going back to (30), we have:

$$F_t - F_{t+1} \geq CF_{t+1}^{3/2} - a_t ,$$

where we denoted $C := \frac{1}{1008\sqrt{M}D^{3/2}}$.

Consequently:

$$\frac{1}{\sqrt{F_{t+1}}} - \frac{1}{\sqrt{F_t}} = \frac{\sqrt{F_t} - \sqrt{F_{t+1}}}{\sqrt{F_tF_{t+1}}}$$

$$\geq \frac{F_t - F_{t+1}}{2F_t\sqrt{F_{t+1}}}.$$

where we used the concavity of $x \mapsto \sqrt{x}$ which gives $\sqrt{a} - \sqrt{b} \leq \frac{1}{2\sqrt{b}}(a - b)$, and took $a = F_{t+1}, b = F_t$.

Thus:

$$\frac{1}{\sqrt{F_{t+1}}} - \frac{1}{\sqrt{F_t}} \geq \frac{CF_{t+1}^{3/2} - a_t}{2F_t\sqrt{F_{t+1}}} = \frac{C}{2}\frac{F_{t+1}}{F_t} - \frac{a_t}{2F_t\sqrt{F_{t+1}}}.$$

Let $\varepsilon = \min_{0\leq t\leq T-1}F_t$. Thus, for all $t \leq T-1$ we have $F_t \geq \varepsilon$, then:

$$\frac{1}{\sqrt{\varepsilon}} \geq \frac{1}{F_T} - \frac{1}{F_0}$$

$$\geq \frac{CT}{2}\left(\frac{1}{T}\sum_{t=0}^{T-1}\frac{F_{t+1}}{F_t}\right) - \frac{T}{2\varepsilon^{3/2}\sqrt{M}}\left(\frac{\sqrt{M}}{T}\sum_{t=0}^{T-1}a_t\right).$$

We note that using the fact that the arithmetic mean is greater than the geometric mean, we have:

$$\frac{1}{T}\sum_{t=0}^{T-1}\frac{F_{t+1}}{F_t} \geq \left(\prod_{t=0}^{T-1}\frac{F_{t+1}}{F_t}\right)^{1/T} = \left(\frac{F_T}{F_0}\right)^{1/T} \geq \left(\frac{\varepsilon}{F_0}\right)^{1/T} \geq 1 + \frac{1}{T}\log\left(\frac{\varepsilon}{F_0}\right),$$

which leads to

$$\frac{1}{\sqrt{\varepsilon}} \geq \frac{CT}{2}\left(1 - \frac{1}{T}\log\left(\frac{F_0}{\varepsilon}\right)\right) - \frac{T}{2\varepsilon^{3/2}\sqrt{M}}\left(\frac{\sqrt{M}}{T}\sum_{t=0}^{T-1}a_t\right).$$

Remember (31), then:

$$\frac{1}{\sqrt{\varepsilon}} \geq \frac{CT}{2}\Big(1 - \frac{1}{T}\log(\frac{F_0}{\varepsilon})\Big) - \frac{T}{2\varepsilon^{3/2}\sqrt{M}}\mathcal{A}(M),$$

where we denoted $\mathcal{A}(M) := 20\frac{a_g^{9/14}\sigma_g^{3/2}}{T^{3/7}} + 803\frac{a_h^{9/5}\sigma_h^3}{M^{3/2}T^{3/5}} + 40\frac{L^{3/8}}{M^{3/8}}\sigma_g^{3/2} + \frac{73(46)^{3/2}L^{3/2}}{M^3}\tilde{\sigma}_h^3(\beta)$. We conclude that:

$$\frac{1}{\sqrt{\varepsilon}} + \frac{C}{2}\log(\frac{F_0}{\varepsilon}) + \frac{T}{2\varepsilon^{3/2}\sqrt{M}}\mathcal{A}(M) \geq \frac{C}{2}T,$$

We invert the previous inequality and get:

$$\varepsilon \leq \frac{36}{C^2T^2} + F_0 e^{-T/3} + \Big(\frac{\mathcal{A}(M)}{C\sqrt{M}}\Big)^{2/3}$$

We conclude that:

$$\min_{0\leq t\leq T-1} F_t \leq \mathcal{O}\Big(\frac{MD^3}{T^2} + \frac{LD\tilde{\sigma}_h^2}{M^2} + \frac{L^{1/4}}{M^{1/4}}D\sigma_g + F_0 e^{-T/3} + \frac{a_g^{3/7}D\sigma_g}{T^{2/7}} + \frac{a_h^{6/5}D\sigma_h^2}{MT^{2/5}}\Big)$$

We choose $M = \max\Big(100L, \frac{L^{1/5}\sigma_g^{4/5}T^{8/5}}{D^{8/5}}, \frac{L^{1/3}\tilde{\sigma}_h^{2/3}}{D^{2/3}}, \frac{a_h^{3/5}\sigma_h T^{4/5}}{D}\Big).$

By replacing M with its value we arrive at the following:

$$\min_{0\leq t\leq T-1} F_t = \mathcal{O}\Big(\frac{LD^3}{T^2} + \frac{L^{1/3}D^{7/3}\tilde{\sigma}_h^{2/3}}{T^{4/3}} + \frac{L^{1/5}D^{7/5}\sigma_g^{4/5}}{T^{2/5}} + F_0 e^{-T/3} + \frac{a_h^{3/5}D^2\sigma_h}{T^{6/5}} + \frac{a_g^{3/7}D\sigma_g}{T^{2/7}}\Big).$$

D   Second-order momentum for the gradients

We will now consider the following momentum:

$$\boldsymbol{g}_t = (1-\alpha)\Big(\boldsymbol{g}_{t-1} + \boldsymbol{H}_t(\boldsymbol{x}_t - \boldsymbol{x}_{t-1})\Big) + \alpha\nabla f_{\xi_t}(\boldsymbol{x}_t), \text{ for } t \geq 1 \text{ and } \boldsymbol{g}_0 = \nabla f_{\xi_0}(\boldsymbol{x}_0), \tag{32}$$

where $\boldsymbol{H}_t$ is the Hessian momentum defined as

$$\boldsymbol{H}_t = (1-\beta)\boldsymbol{H}_{t-1} + \beta\nabla^2 f_{\xi_t}(\boldsymbol{x}_t), \text{ for } t \geq 1 \text{ and } \boldsymbol{g}_0 = \nabla^2 f_{\xi_0}(\boldsymbol{x}_0) , \tag{33}$$

for $\alpha, \beta \in (0, 1]$.

Lemma D.1. *For the update in (32), we have for any $\gamma > 0$:*

$$\frac{1}{2T}\sum_{t=0}^{T-1}\mathbb{E}\|\varepsilon_t\|^{3/2} \leq \frac{\sigma_{g,0}^{3/2}}{\alpha T} + \frac{(1-\beta)\sigma_{h,0}^3}{\gamma^{3/2}\alpha^{3/2}\beta T} + \alpha^{3/4}\sigma_g^{3/2} + \frac{\beta^{3/2}}{\gamma^{3/2}\alpha^{3/2}}\tilde{\sigma}_h^3(\beta)$$

$$+ (1-\alpha)\Big(\frac{L^{3/2}}{\alpha^{3/2}} + \frac{\gamma^{3/2}}{\alpha^{3/2}} + \frac{1-\beta}{2^{3/2}\gamma^{3/2}\alpha^{3/2}\beta^3}\Big)\frac{1}{T}\sum_{t=0}^{T-1}\mathbb{E}r_t^3.$$

Proof. As before, let $\varepsilon_t = \boldsymbol{g}_t - \nabla f(\boldsymbol{x}_t)$, then we have:

$$\varepsilon_t = (1-\alpha)\Big(\boldsymbol{g}_{t-1} + \boldsymbol{H}_t(\boldsymbol{x}_t - \boldsymbol{x}_{t-1})\Big) + \alpha\nabla f_{\xi_t}(\boldsymbol{x}_t) - \nabla f(\boldsymbol{x}_t)$$

$$= (1-\alpha)\Big(\varepsilon_{t-1} + \nabla f(\boldsymbol{x}_{t-1}) - \nabla f(\boldsymbol{x}_t) + \boldsymbol{H}_t(\boldsymbol{x}_t - \boldsymbol{x}_{t-1})\Big) + \alpha(\nabla f_{\xi_t}(\boldsymbol{x}_t) - \nabla f(\boldsymbol{x}_t))$$

$$= (1-\alpha)\Big(\varepsilon_{t-1} + \boldsymbol{z}(\boldsymbol{x}_{t-1}, \boldsymbol{x}_t) + [\boldsymbol{H}_t - \nabla^2 f(\boldsymbol{x}_t)](\boldsymbol{x}_t - \boldsymbol{x}_{t-1})\Big) + \alpha(\nabla f_{\xi_t}(\boldsymbol{x}_t) - \nabla f(\boldsymbol{x}_t))$$

$$= (1-\alpha)\Big(\varepsilon_{t-1} + \boldsymbol{z}(\boldsymbol{x}_{t-1}, \boldsymbol{x}_t) + \boldsymbol{\Sigma}_t(\boldsymbol{x}_t - \boldsymbol{x}_{t-1})\Big) + \alpha\boldsymbol{n}_t.$$

Thus

$$\boldsymbol{\varepsilon}_t = (1-\alpha)^t \boldsymbol{\varepsilon}_0 + (1-\alpha) \sum_{i=0}^{t-1} (1-\alpha)^i \Big(\boldsymbol{z}(\boldsymbol{x}_{t-i-1}, \boldsymbol{x}_{t-i}) + \boldsymbol{\Sigma}_{t-i}(\boldsymbol{x}_{t-i} - \boldsymbol{x}_{t-i-1}) \Big) + \alpha \sum_{i=0}^{t-1} (1-\alpha)^i \boldsymbol{n}_{t-i}.$$

By taking the norm, we get:

$$\|\boldsymbol{\varepsilon}_t\| \leq (1-\alpha)^t \|\boldsymbol{n}_0\| + (1-\alpha) \sum_{i=0}^{t-1} (1-\alpha)^i \Big(\frac{L}{2} r_{t-i}^2 + \|\boldsymbol{\Sigma}_{t-i}\| r_{t-i} \Big) + \alpha \| \sum_{i=0}^{t-1} (1-\alpha)^i \boldsymbol{n}_{t-i}\|$$

$$\leq (1-\alpha)^t \|\boldsymbol{n}_0\| + (1-\alpha) \frac{L+\gamma}{2} \sum_{i=0}^{t-1} (1-\alpha)^i r_{t-i}^2 + (1-\alpha) \frac{1}{2\gamma} \sum_{i=0}^{t-1} (1-\alpha)^i \|\boldsymbol{\Sigma}_{t-i}\|^2 + \alpha \| \sum_{i=0}^{t-1} (1-\alpha)^i \boldsymbol{n}_{t-i}\|,$$

where $\gamma > 0$ is to be chosen latter. Now we take the power $3/2$ and get:

$$\frac{1}{2} \|\boldsymbol{\varepsilon}_t\|^{3/2} \leq (1-\alpha)^t \|\boldsymbol{n}_0\|^{3/2} + (1-\alpha) \Big(\frac{L+\gamma}{2} \sum_{i=0}^{t-1} (1-\alpha)^i r_{t-i}^2 \Big)^{3/2} + (1-\alpha) \Big(\frac{1}{2\gamma} \sum_{i=0}^{t-1} (1-\alpha)^i \|\boldsymbol{\Sigma}_{t-i}\|^2 \Big)^{3/2}$$

$$+ \alpha^{3/2} \| \sum_{i=0}^{t-1} (1-\alpha)^i \boldsymbol{n}_{t-i}\|^{3/2}$$

$$\leq (1-\alpha)^t \|\boldsymbol{n}_0\|^{3/2} + (1-\alpha) \frac{(L+\gamma)^{3/2}}{2^{3/2} \alpha^{1/2}} \sum_{i=0}^{t-1} (1-\alpha)^i r_{t-i}^3 + (1-\alpha) \frac{1}{2^{3/2} \gamma^{3/2} \alpha^{1/2}} \sum_{i=0}^{t-1} (1-\alpha)^i \|\boldsymbol{\Sigma}_{t-i}\|^3$$

$$+ \alpha^{3/2} \| \sum_{i=0}^{t-1} (1-\alpha)^i \boldsymbol{n}_{t-i}\|^{3/2}.$$

By taking the expectation, and using the same reasoning as before, we get:

$$\frac{1}{2} \mathbb{E}\|\boldsymbol{\varepsilon}_t\|^{3/2} \leq (1-\alpha)^t \sigma_{g,0}^{3/2} + (1-\alpha) \frac{(L+\gamma)^{3/2}}{2^{3/2} \alpha^{1/2}} \sum_{i=0}^{t-1} (1-\alpha)^i \mathbb{E}r_{t-i}^3 + \alpha^{3/4} \sigma_g^{3/2} + (1-\alpha) \frac{1}{2^{3/2} \gamma^{3/2} \alpha^{1/2}} \sum_{i=0}^{t-1} (1-\alpha)^i \mathbb{E}\|\boldsymbol{\Sigma}_{t-i}\|^3.$$

By substituting our bound of $\mathbb{E}\|\boldsymbol{\Sigma}_{t-i}\|^3$ from before, we get:

$$\frac{1}{9} \sum_{i=0}^{t-1} (1-\alpha)^i \mathbb{E}\|\boldsymbol{\Sigma}_{t-i}\|^3 \leq \sum_{i=0}^{t-1} (1-\alpha)^i \Big((1-\beta)^{t-i} \sigma_{h,0}^3 + \beta^{3/2} \tilde{\sigma}_h^3(\beta) + \frac{(1-\beta)L^3}{\beta^2} \sum_{j=0}^{t-i-1} (1-\beta)^j \mathbb{E}r_{t-i-j}^3 \Big)$$

$$\leq \sum_{i=0}^{t-1} (1-\alpha)^i \Big((1-\beta)^{t-i} \sigma_{h,0}^3 + \beta^{3/2} \tilde{\sigma}_h^3(\beta) + \frac{(1-\beta)L^3}{\beta^2} \sum_{j=0}^{t-i-1} (1-\beta)^j \mathbb{E}r_{t-i-j}^3 \Big)$$

$$\leq (1-\beta) \frac{(1-\beta)^t - (1-\alpha)^t}{\alpha - \beta} \sigma_{h,0}^3 + \frac{\beta^{3/2}}{\alpha} \tilde{\sigma}_h^3(\beta) + \frac{(1-\beta)L^3}{\beta^2} \sum_{i=0}^{t-1} (1-\alpha)^i \sum_{j=0}^{t-i-1} (1-\beta)^j \mathbb{E}r_{t-i-j}^3$$

$$\leq (1-\beta) \frac{(1-\min(\alpha,\beta))^t}{\max(\alpha,\beta)} \sigma_{h,0}^3 + \frac{\beta^{3/2}}{\alpha} \tilde{\sigma}_h^3(\beta) + \frac{(1-\beta)L^3}{\beta^2} \sum_{i=0}^{t-1} \frac{(1-\min(\alpha,\beta))^i}{\max(\alpha,\beta)} \mathbb{E}r_{t-i}^3.$$

Thus,

$$\frac{1}{2T} \sum_{t=0}^{T-1} \mathbb{E}\|\boldsymbol{\varepsilon}_t\|^{3/2} \leq \frac{\sigma_{g,0}^{3/2}}{\alpha T} + \frac{(1-\beta)\sigma_{h,0}^3}{\gamma^{3/2} \alpha^{3/2} \beta T} + \alpha^{3/4} \sigma_g^{3/2} + \frac{\beta^{3/2}}{\gamma^{3/2} \alpha^{3/2}} \tilde{\sigma}_h^3(\beta)$$

$$+ (1-\alpha) \Big(\frac{L^{3/2}}{\alpha^{3/2}} + \frac{\gamma^{3/2}}{\alpha^{3/2}} + \frac{1-\beta}{2^{3/2} \gamma^{3/2} \alpha^{3/2} \beta^3} \Big) \frac{1}{T} \sum_{t=0}^{T-1} \mathbb{E}r_t^3.$$

$\square$

This means we need to choose α, β such that

$$\frac{L^{3/2}}{\alpha^{3/2}} + \frac{\gamma^{3/2}}{\alpha^{3/2}} + \frac{(1-\beta)L^3}{2^{3/2}\gamma^{3/2}\alpha^{3/2}\beta^3} + (1-\beta)\frac{73}{M^{3/2}}\frac{9L^3}{\beta^3} - \frac{M^{3/2}}{72} \leq 0.$$

For example, for $\beta = 1$ (smaller choices of β will make the middle term very high and are less likely to lead to any improvement), it suffices to take: $\alpha = 3175 \max(\frac{L}{M}, \frac{\gamma}{M})$.

We choose γ such that it minimizes $\frac{\gamma^{3/4}}{M^{3/4}}\sigma_g^{3/2} + \frac{1}{\gamma^{3/2}\alpha^{3/2}}\sigma_h^3$ and $\gamma \leq M/3175$ to ensure α can be made ≤ 1. We choose $\gamma = \min(M, \frac{M^{3/5}\sigma_h^{4/5}}{\sigma_g^{2/5}})$. Note that when $\gamma = M/3175$, we have $\sigma_g \leq 3175^{5/2}\frac{\sigma_h^2}{M}$.

For this choice of γ we get: $\frac{\gamma^{3/4}}{M^{3/4}}\sigma_g^{3/2} + \frac{1}{\gamma^{3/2}\alpha^{3/2}}\sigma_h^3 \leq \frac{2\sigma_h^3}{M^{3/2}} + \frac{2\sigma_g^{9/10}\sigma_h^{3/5}}{M^{3/10}}$.

We also choose $M \geq 3175L$ to ensure $\alpha \leq 1$.

Thus

$$\frac{1}{1008}\frac{1}{T}\sum_{t=0}^{T-1}\mathbb{E}\mu_M(\boldsymbol{x}_{t+1}) \leq \frac{\sqrt{M}F_0}{T} - \frac{M^{3/2}}{72}\mathbb{E}\frac{1}{T}\sum_{t=0}^{T-1}r_t^3 + 4\frac{1}{T}\sum_{t=0}^{T-1}\mathbb{E}\|\boldsymbol{\varepsilon}_t\|^{3/2} + \frac{73}{M^{3/2}}\frac{1}{T}\sum_{t=0}^{T-1}\mathbb{E}\|\boldsymbol{\Sigma}_t\|^3$$

$$\leq \frac{\sqrt{M}F_0}{T} + \frac{2}{\alpha T}\sigma_{g,0}^{3/2} + 2\alpha^{3/4}\sigma_g^{3/2} + 2\frac{1}{\gamma^{3/2}\alpha^{3/2}}\sigma_h^3 + \frac{73}{M^{3/2}}\sigma_h^3$$

$$= \mathcal{O}\Big(\frac{\sqrt{M}F_0}{T} + \frac{M}{LT}\sigma_{g,0}^{3/2} + \frac{\sigma_g^{9/10}\sigma_h^{3/5}}{M^{3/10}} + \frac{L^{3/4}}{M^{3/4}}\sigma_g^{3/2} + \frac{1}{M^{3/2}}\sigma_h^3\Big).$$

We see that we have a term $\frac{\sigma_g^{9/10}\sigma_h^{3/5}}{M^{3/10}}$ that is much slower (for large M) than the $\frac{L^{3/8}}{M^{3/8}}\sigma_g^{3/2}$ we had before.

By sampling an initial batch $T/100$ larger than other batches, we ensure $\sigma_{g,0}^{3/2} = \mathcal{O}(\frac{\sigma_g^{3/2}}{T^{3/4}})$.

We can optimize the right-hand side over M as before, but we can already see that the obtained rate will worsen. Note also that the Hessian term $\frac{1}{M^{3/2}}\sigma_h^3$ is smaller than the one we had before (since we needed to set $\beta = 1$).

To be specific, by choosing $M = \max(3157L, \frac{\sigma_g^{9/8}\sigma_h^{3/4}T^{5/4}}{F_0^{5/4}}, \frac{\sigma_g^{6/5}L^{3/5}T^{4/5}}{F_0^{4/5}}, \frac{\sigma_h^2 T^{1/2}}{F_0^{1/2}})$, we get:

$$\frac{1}{T}\sum_{t=0}^{T-1}\mathbb{E}\mu_M(\boldsymbol{x}_{t+1}) = \mathcal{O}\Big(\frac{\sqrt{L}F_0}{T} + \frac{\sigma_g^{9/16}\sigma_h^{3/8}F_0^{3/8}}{T^{3/8}} + \frac{L^{3/10}\sigma_g^{3/5}F_0^{3/5}}{T^{3/5}} + \frac{\sigma_h^{3/4}F_0^{3/4}}{T^{3/4}} + \frac{M}{LT^{7/4}}\sigma_g^{3/2}\Big)$$

The term $\frac{M}{LT^{7/4}}\sigma_g^{3/2}$ leads to smaller terms compared to the ones we already have.

E Analyzing the HB momentum for gradient and Hessian.

We consider the Heavy-Ball (HB) momentum defined by the recursive relation:

$$\boldsymbol{g}_t = (1-\alpha)\boldsymbol{g}_{t-1} + \alpha\nabla_{\xi_t}f(\boldsymbol{x}_t) . \tag{34}$$

To analyze this momentum, we need the following (additional) assumption:

Assumption E.1. *The gradients of the function f are L_g-Lipschitz for some constant $L_g \geq 0$, meaning that for all $\boldsymbol{x}, \boldsymbol{y}$ we have*
$$\|\nabla f(\boldsymbol{x}) - \nabla f(\boldsymbol{y})\| \leq L_g\|\boldsymbol{x} - \boldsymbol{y}\| .$$

We prove the following Lemma:

Lemma E.2. *Under Assumption E.1, for any $\gamma > 0$:*

$$\frac{1}{T}\sum_{t=0}^{T-1}\mathbb{E}\|\varepsilon_t\|^{3/2} \leq \sqrt{3}\left(\frac{1}{\alpha T}\sigma_{g,0}^{3/2} + \alpha^{3/4}\sigma_g^{3/2} + (1-\alpha)^{3/2}\frac{L_g^{3/2}}{\alpha^{3/2}\gamma^{3/2}} + (1-\alpha)^{3/2}\frac{L_g^{3/2}\gamma^{3/2}}{\alpha^{3/2}}\frac{1}{T}\sum_{i=0}^{t-1}\mathbb{E}r_t^3\right)$$

Proof. For this momentum, we have:

$$\begin{aligned}
\varepsilon_t &= (1-\alpha)\boldsymbol{g}_{t-1} + \alpha\nabla f_{\xi_t}(\boldsymbol{x}_t) - \nabla f(\boldsymbol{x}_t) \\
&= (1-\alpha)\big(\boldsymbol{g}_{t-1} - \nabla f(\boldsymbol{x}_t)\big) + \alpha\underbrace{\big(\nabla f_{\xi_t}(\boldsymbol{x}_t) - \nabla f(\boldsymbol{x}_t)\big)}_{:=\boldsymbol{n}_t} \\
&= (1-\alpha)\big(\varepsilon_{t-1} + \underbrace{\nabla f(\boldsymbol{x}_{t-1}) - \nabla f(\boldsymbol{x}_t)}_{:=\boldsymbol{z}(\boldsymbol{x}_{t-1},\boldsymbol{x}_t)}\big) + \alpha\boldsymbol{n}_t \\
&= (1-\alpha)\big(\varepsilon_{t-1} + \boldsymbol{z}(\boldsymbol{x}_{t-1},\boldsymbol{x}_t)\big) + \alpha\boldsymbol{n}_t,
\end{aligned}$$

where we defined $\boldsymbol{z}(\boldsymbol{a},\boldsymbol{b}) := \nabla f(\boldsymbol{a}) - \nabla f(\boldsymbol{b})$. According to Assumption E.1: $\|\boldsymbol{z}(\boldsymbol{a},\boldsymbol{b})\| \leq L_g\|\boldsymbol{a}-\boldsymbol{b}\|$.

Thus

$$\varepsilon_t = (1-\alpha)^t\varepsilon_0 + (1-\alpha)\sum_{i=0}^{t-1}(1-\alpha)^i\boldsymbol{z}(\boldsymbol{x}_{t-i-1},\boldsymbol{x}_{t-i}) + \alpha\sum_{i=0}^{t-1}(1-\alpha)^i\boldsymbol{n}_{t-i}\ ,\ t \geq 1.$$

And

$$\|\varepsilon_t\| \leq (1-\alpha)^t\|\varepsilon_0\| + (1-\alpha)L_g\sum_{i=0}^{t-1}(1-\alpha)^i r_{t-i} + \alpha\|\sum_{i=0}^{t-1}(1-\alpha)^i\boldsymbol{n}_{t-i}\|.$$

Now we raise this to the power $3/2$ and using the convexity of $x \mapsto x^{3/2}$, we get:

$$\begin{aligned}
\|\varepsilon_t\|^{3/2} &\leq \sqrt{3}\left((1-\alpha)^{3t/2}\|\varepsilon_0\|^{3/2} + (1-\alpha)^{3/2}L_g^{3/2}\Big(\sum_{i=0}^{t-1}(1-\alpha)^i r_{t-i}\Big)^{3/2} + \alpha^{3/2}\|\sum_{i=0}^{t-1}(1-\alpha)^i\boldsymbol{n}_{t-i}\|^{3/2}\right) \\
&\leq \sqrt{3}\left((1-\alpha)^{3t/2}\|\varepsilon_0\|^{3/2} + (1-\alpha)^{3/2}L_g^{3/2}\Big(\sum_{i=0}^{t-1}(1-\alpha)^i r_{t-i}\Big)^{3/2} + \alpha^{3/2}\|\sum_{i=0}^{t-1}(1-\alpha)^i\boldsymbol{n}_{t-i}\|^{3/2}\right) \\
&\leq \sqrt{3}\left((1-\alpha)^{3t/2}\|\varepsilon_0\|^{3/2} + (1-\alpha)^{3/2}\frac{L_g^{3/2}}{\sqrt{\alpha}}\sum_{i=0}^{t-1}(1-\alpha)^i r_{t-i}^{3/2} + \alpha^{3/2}\|\sum_{i=0}^{t-1}(1-\alpha)^i\boldsymbol{n}_{t-i}\|^{3/2}\right).
\end{aligned}$$

Now we take the expectation, and in the same lines as for the (IT) momentum, we get:

$$\mathbb{E}\|\varepsilon_t\|^{3/2} \leq \sqrt{3}\left((1-\alpha)^{3t/2}\sigma_{g,0}^{3/2} + (1-\alpha)^{3/2}\frac{L_g^{3/2}}{\sqrt{\alpha}}\sum_{i=0}^{t-1}(1-\alpha)^i\mathbb{E}r_{t-i}^{3/2} + \alpha^{3/4}\sigma_g^{3/2}\right).$$

We take now the average over iterations:

$$\frac{1}{T}\sum_{t=0}^{T-1}\mathbb{E}\|\varepsilon_t\|^{3/2} \leq \sqrt{3}\left(\frac{1}{\alpha T}\sigma_{g,0}^{3/2} + (1-\alpha)^{3/2}\frac{L_g^{3/2}}{\alpha^{3/2}}\frac{1}{T}\sum_{t=0}^{t-1}\mathbb{E}r_t^{3/2} + \alpha^{3/4}\sigma_g^{3/2}\right).$$

Then we use Young's inequality to transform the $r^{3/2}$ terms to r^3 terms, we get for any $\gamma > 0$:

$$\frac{1}{T}\sum_{t=0}^{T-1}\mathbb{E}\|\varepsilon_t\|^{3/2} \leq \sqrt{3}\left(\frac{1}{\alpha T}\sigma_{g,0}^{3/2} + \alpha^{3/4}\sigma_g^{3/2} + (1-\alpha)^{3/2}\frac{L_g^{3/2}}{\alpha^{3/2}\gamma^{3/2}} + (1-\alpha)^{3/2}\frac{L_g^{3/2}\gamma^{3/2}}{\alpha^{3/2}}\frac{1}{T}\sum_{i=0}^{t-1}\mathbb{E}r_t^3\right).$$

This finishes the proof. □

Convergence analysis.

> **Theorem E.3.** *Using Cubic Newton with the gradient and Hessian updates in (34) and (24) respectively with $\alpha = \min(1, \frac{18L_g^{4/5}}{M^{2/5}\sigma_g^{2/5}}), \beta = 1$ and under the additional Assumption E.1, is such that for any $M \geq L$ we have*
>
> $$\frac{1}{1008}\mathbb{E}\mu_M(\boldsymbol{x}_{out}^T) \leq \frac{\sqrt{M}\tilde{F}_{0M}}{T} + 71\frac{L_g^{3/5}\sigma_g^{6/5}}{M^{3/10}} + \frac{\tilde{\sigma}_h^3}{M^{3/2}}$$
>
> *where $\tilde{F}_{0M} = F_0 + \frac{12\sigma_g^{2/5}}{18L_g^{4/5}M^{1/10}}\sigma_{g,0}^{3/2} \leq \tilde{F}_{0L} := F_0 + \frac{12\sigma_g^{2/5}}{18L_g^{4/5}L^{1/10}}\sigma_{g,0}^{3/2}$ and $\tilde{\sigma}_h^3 = 12^{15/4}L_g^3 + 73\sigma_h^3.$, the last inequality is obtained because $M \leq L$.*

Proof. We consider $\beta = 1$ (other values do not improve the rate), Applying Theorem B.1

$$\frac{1}{1008}\frac{1}{T}\sum_{t=0}^{T-1}\mathbb{E}\mu_M(\boldsymbol{x}_{t+1}) \leq \frac{\sqrt{M}F_0}{T} - \frac{M^{3/2}}{72}\mathbb{E}\frac{1}{T}\sum_{t=0}^{T-1}r_t^3 + 4\frac{1}{T}\sum_{t=0}^{T-1}\mathbb{E}\|\boldsymbol{\varepsilon}_t\|^{3/2} + \frac{73}{M^{3/2}}\frac{1}{T}\sum_{t=0}^{T-1}\mathbb{E}\|\boldsymbol{\Sigma}_t\|^3$$

$$\leq \frac{\sqrt{M}F_0}{T} + \frac{12}{\alpha T}\sigma_{g,0}^{3/2} + 8\alpha^{3/4}\sigma_g^{3/2} + 8(1-\alpha)^{3/2}\frac{L_g^{3/2}}{\alpha^{3/2}\gamma^{3/2}} + \frac{73}{M^{3/2}}\sigma_h^3$$

$$+ \left((1-\alpha)^{3/2}\frac{L_g^{3/2}\gamma^{3/2}}{\alpha^{3/2}} - \frac{M^{3/2}}{72}\right)\mathbb{E}\frac{1}{T}\sum_{t=0}^{T-1}r_t^3.$$

We choose α such that $(1-\alpha)^{3/2}\frac{L_g^{3/2}\gamma^{3/2}}{\alpha^{3/2}} - \frac{M^{3/2}}{72} \leq 0$, for example $\alpha = 18\frac{L_g\gamma}{M}$, thus

$$\frac{1}{1008}\frac{1}{T}\sum_{t=0}^{T-1}\mathbb{E}\mu_M(\boldsymbol{x}_{t+1}) \leq \frac{\sqrt{M}F_0}{T} + \frac{12}{\alpha T}\sigma_{g,0}^{3/2} + 70(\frac{L_g\gamma}{M})^{3/4}\sigma_g^{3/2} + 0.1\frac{M^{3/2}}{\gamma^3} + \frac{73}{M^{3/2}}\sigma_h^3.$$

We choose $\gamma \in \arg\min_{\gamma \in (0, \frac{M}{18L_g}]}70(\frac{L_g\gamma}{M})^{3/4}\sigma_g^{3/2} + 0.1\frac{M^{3/2}}{\gamma^3}$, we find $\gamma = \min(\frac{M}{18L_g}, \frac{M^{3/5}}{L_g^{1/5}\sigma_g^{2/5}})$. Therefore, $\alpha = \min(1, \frac{18L_g^{4/5}}{M^{2/5}\sigma_g^{2/5}})$ and we get:

$$\frac{1}{1008}\frac{1}{T}\sum_{t=0}^{T-1}\mathbb{E}\mu_M(\boldsymbol{x}_{t+1}) \leq \frac{\sqrt{M}F_0}{T} + \frac{12M^{2/5}\sigma_g^{2/5}}{18L_g^{4/5}T}\sigma_{g,0}^{3/2} + 71\frac{L_g^{3/5}\sigma_g^{6/5}}{M^{3/10}} + 12^{15/4}\frac{L_g^3}{M^{3/2}} + \frac{73}{M^{3/2}}\sigma_h^3.$$

The term $\frac{L_g^3}{M^{3/2}}$ comes from the constraint $\gamma \leq \frac{M}{18L_g}$ being activated which means $\sigma_g^{3/2} \leq 12^{15/4}\frac{L_g^3}{M^{3/2}}$. This term limits any possible improvement of the Hessian term (since it has the same power of M), hence why we chose $\beta = 1$ from the start.

Finally, we obtain

$$\frac{1}{1008}\frac{1}{T}\sum_{t=0}^{T-1}\mathbb{E}\mu_M(\boldsymbol{x}_{t+1}) \leq \frac{\sqrt{M}\tilde{F}_{0M}}{T} + 71\frac{L_g^{3/5}\sigma_g^{6/5}}{M^{3/10}} + \frac{12^{15/4}L_g^3 + 73\sigma_h^3}{M^{3/2}},$$

where $\tilde{F}_{0M} = F_0 + \frac{12\sigma_g^{2/5}}{18L_g^{4/5}M^{1/10}}\sigma_{g,0}^{3/2} \leq \tilde{F}_{0L} := F_0 + \frac{12\sigma_g^{2/5}}{18L_g^{4/5}L^{1/10}}\sigma_{g,0}^{3/2}$, the last inequality is obtained because $M \leq L$.

$\square$

Choosing $M = \max(L, \frac{L_g^{3/4}\sigma_g^{3/2}T^{5/4}}{\tilde{F}_{0L}^{5/4}}, \frac{\tilde{\sigma}_h^2 T^{1/2}}{\tilde{F}_{0L}^{1/2}})$, we get:

$$\frac{1}{1008}\frac{1}{T}\sum_{t=0}^{T-1}\mathbb{E}\mu_M(\boldsymbol{x}_{t+1}) \leq \frac{\sqrt{L}\tilde{F}_{0L}}{T} + 72\left(\frac{L_g\tilde{F}_{0L}\sigma_g^2}{T}\right)^{3/8} + \left(\frac{\tilde{F}_{0L}\tilde{\sigma}_h}{T}\right)^{3/4}$$

F Analyzing the MVR momentum for the gradient.

We consider the Heavy-Ball (MVR) momentum defined by the recursive relation:

$$\boldsymbol{g}_t = (1-\alpha)\Big(\boldsymbol{g}_{t-1} + \nabla_{\xi_t} f(\boldsymbol{x}_t) - \nabla_{\xi_t} f(\boldsymbol{x}_{t-1})\Big) + \alpha \nabla_{\xi_t} f(\boldsymbol{x}_t) . \tag{35}$$

To analyze this momentum, we need a stronger first-order smoothness assumption:

Assumption F.1. *There exists $L_g \geq 0$ such that for all ξ, the gradients of the function f_ξ are L_g-Lipschitz, meaning that for all $\xi, \boldsymbol{x}, \boldsymbol{y}$ we have*

$$\|\nabla f_\xi(\boldsymbol{x}) - \nabla f_\xi(\boldsymbol{y})\| \leq L_g \|\boldsymbol{x} - \boldsymbol{y}\| .$$

We prove the following Lemma:

Lemma F.2. *For the gradient update defined in (35), we have for any $\gamma > 0$:*

$$\frac{1}{T}\sum_{t=0}^{T-1}\mathbb{E}\|\varepsilon_t\|^{3/2} \leq \frac{1}{\alpha T}\sigma_{g,0}^{3/2} + \alpha^{3/2}\sigma_g^{3/2} + \frac{8(1-\alpha)^{3/2}L_g^{3/2}}{\alpha\gamma^{3/2}} + \frac{8(1-\alpha)^{3/2}L_g^{3/2}\gamma^{3/2}}{\alpha}\frac{1}{T}\sum_{i=0}^{t-1}r_t^3$$

Proof. For this momentum, we have:

$$\varepsilon_t = (1-\alpha)\Big(\boldsymbol{g}_{t-1} + \nabla_{\xi_t} f(\boldsymbol{x}_t) - \nabla_{\xi_t} f(\boldsymbol{x}_{t-1})\Big) + \alpha \nabla f_{\xi_t}(\boldsymbol{x}_t) - \nabla f(\boldsymbol{x}_t)$$

$$= (1-\alpha)\Big(\boldsymbol{g}_{t-1} + \nabla_{\xi_t} f(\boldsymbol{x}_t) - \nabla_{\xi_t} f(\boldsymbol{x}_{t-1}) - \nabla f(\boldsymbol{x}_t)\Big) + \alpha \underbrace{\Big(\nabla f_{\xi_t}(\boldsymbol{x}_t) - \nabla f(\boldsymbol{x}_t)\Big)}_{:=\boldsymbol{n}_t}$$

$$= (1-\alpha)\Big(\varepsilon_{t-1} + \underbrace{\nabla f(\boldsymbol{x}_{t-1}) - \nabla f(\boldsymbol{x}_t) + \nabla_{\xi_t} f(\boldsymbol{x}_t) - \nabla_{\xi_t} f(\boldsymbol{x}_{t-1})}_{:=\boldsymbol{z}_{\xi_t}(\boldsymbol{x}_{t-1},\boldsymbol{x}_t)}\Big) + \alpha \boldsymbol{n}_t$$

$$= (1-\alpha)\Big(\varepsilon_{t-1} + \boldsymbol{z}_{\xi_t}(\boldsymbol{x}_{t-1},\boldsymbol{x}_t)\Big) + \alpha \boldsymbol{n}_t,$$

where we defined $\boldsymbol{z}_\xi(\boldsymbol{a},\boldsymbol{b}) := \nabla f(\boldsymbol{a}) - \nabla f(\boldsymbol{b}) + \nabla f_\xi(\boldsymbol{b}) - \nabla f_\xi(\boldsymbol{a})$. According to Assumption F.1: $\|\boldsymbol{z}_\xi(\boldsymbol{a},\boldsymbol{b})\| \leq 2L_g\|\boldsymbol{a}-\boldsymbol{b}\|$.

Note that knowing $\boldsymbol{x}_{t-1}, \boldsymbol{x}_t$, the random variable $(1-\alpha)\boldsymbol{z}_{\xi_t}(\boldsymbol{x}_{t-1},\boldsymbol{x}_t) + \alpha\boldsymbol{n}_t$ is centred and independent from ε_{t-1}.

Thus

$$\mathbb{E}\|\varepsilon_t\|^2 = (1-\alpha)^2\mathbb{E}\|\varepsilon_{t-1}\|^2 + \mathbb{E}\|(1-\alpha)\boldsymbol{z}_{\xi_t}(\boldsymbol{x}_{t-1},\boldsymbol{x}_t) + \alpha\boldsymbol{n}_t\|^2$$

$$\leq (1-\alpha)^2\mathbb{E}\|\varepsilon_{t-1}\|^2 + 8(1-\alpha)^2 L_g^2 r_t^2 + 2\alpha^2\sigma_g^2$$

$$\leq (1-\alpha)^{2t}\|\varepsilon_0\|^2 + 8(1-\alpha)^2 L_g^2\sum_{i=0}^{t-1}(1-\alpha)^{2i}r_{t-i}^2 + \alpha^2\sum_{i=0}^{t-1}(1-\alpha)^{2i}\sigma_g^2$$

$$\leq (1-\alpha)^{2t}\sigma_{g,0}^2 + 8(1-\alpha)^2 L_g^2\sum_{i=0}^{t-1}(1-\alpha)^{2i}r_{t-i}^2 + \alpha\sigma_g^2.$$

Using Jensen's inequality, we get:

$$\mathbb{E}\|\varepsilon_t\|^{3/2} \leq (1-\alpha)^{3t/2}\sigma_{g,0}^{3/2} + 8(1-\alpha)^{3/2}L_g^{3/2}\Big(\sum_{i=0}^{t-1}(1-\alpha)^{2i}r_{t-i}^2\Big)^{3/4} + \alpha^{3/2}\sigma_g^{3/2}$$

$$\leq (1-\alpha)^{3t/2}\sigma_{g,0}^{3/2} + 8(1-\alpha)^{3/2}L_g^{3/2}\sum_{i=0}^{t-1}(1-\alpha)^{3i/2}r_{t-i}^{3/2} + \alpha^{3/2}\sigma_g^{3/2}.$$

We take the average and get:

$$\frac{1}{T}\sum_{t=0}^{T-1}\mathbb{E}\|\varepsilon_t\|^{3/2} \le \frac{1}{\alpha T}\sigma_{g,0}^{3/2} + \alpha^{3/2}\sigma_g^{3/2} + \frac{8(1-\alpha)^{3/2}L_g^{3/2}}{\alpha}\frac{1}{T}\sum_{i=0}^{t-1}r_t^{3/2}$$

$$\le \frac{1}{\alpha T}\sigma_{g,0}^{3/2} + \alpha^{3/2}\sigma_g^{3/2} + \frac{8(1-\alpha)^{3/2}L_g^{3/2}}{\alpha\gamma^{3/2}} + \frac{8(1-\alpha)^{3/2}L_g^{3/2}\gamma^{3/2}}{\alpha}\frac{1}{T}\sum_{i=0}^{t-1}r_t^3,$$

for any $\gamma > 0$. This finishes the proof.

$\square$

Theorem F.3. *Using Cubic Newton with the gradient and Hessian updates in (35) and (24) respectively with $\alpha = \min(1, \frac{18L_g^{4/5}}{M^{2/5}\sigma_g^{2/5}}), \beta = 1$ and under the additional Assumption F.1, is such that for any $M \ge L$ we have*

$$\frac{1}{1008}\mathbb{E}\mu_M(\boldsymbol{x}_{out}^T) \le \frac{\sqrt{M}\tilde{F}_{0M}}{T} + \frac{M^{6/11}\sigma_g^{6/11}}{192L_g^{12/11}T}\sigma_{g,0}^{3/2} + 327\frac{L_g^{9/11}\sigma_g^{12/11}}{M^{9/22}} + \frac{\tilde{\sigma}_h^3}{M^{3/2}},$$

where $\tilde{F}_{0M} = F_0 + \frac{M^{1/22}\sigma_g^{6/11}}{192L_g^{12/11}}\sigma_{g,0}^{3/2}$. and $\tilde{\sigma}_h^3 = (72 \times 32)^{11/6}L_g^3 + 73\sigma_h^3$.

Proof. We consider $\beta = 1$ (other values do not improve the rate), Applying Theorem B.1

$$\frac{1}{1008}\frac{1}{T}\sum_{t=0}^{T-1}\mathbb{E}\mu_M(\boldsymbol{x}_{t+1}) \le \frac{\sqrt{M}F_0}{T} - \frac{M^{3/2}}{72}\mathbb{E}\frac{1}{T}\sum_{t=0}^{T-1}r_t^3 + 4\frac{1}{T}\sum_{t=0}^{T-1}\mathbb{E}\|\varepsilon_t\|^{3/2} + \frac{73}{M^{3/2}}\frac{1}{T}\sum_{t=0}^{T-1}\mathbb{E}\|\Sigma_t\|^3$$

$$\le \frac{\sqrt{M}F_0}{T} + \frac{12}{\alpha T}\sigma_{g,0}^{3/2} + 32\alpha^{3/4}\sigma_g^{3/2} + 32(1-\alpha)^{3/2}\frac{L_g^{3/2}}{\alpha\gamma^{3/2}} + \frac{73}{M^{3/2}}\sigma_h^3$$

$$+ \left(32(1-\alpha)^{3/2}\frac{L_g^{3/2}\gamma^{3/2}}{\alpha} - \frac{M^{3/2}}{72}\right)\mathbb{E}\frac{1}{T}\sum_{t=0}^{T-1}r_t^3.$$

We choose α such that $32\frac{L_g^{3/2}\gamma^{3/2}}{\alpha} - \frac{M^{3/2}}{72} \le 0$, for example $\alpha = 72 \times 32\left(\frac{L_g\gamma}{M}\right)^{3/2}$, thus

$$\frac{1}{1008}\frac{1}{T}\sum_{t=0}^{T-1}\mathbb{E}\mu_M(\boldsymbol{x}_{t+1}) \le \frac{\sqrt{M}F_0}{T} + \frac{12}{\alpha T}\sigma_{g,0}^{3/2} + 326(\frac{L_g\gamma}{M})^{9/8}\sigma_g^{3/2} + \frac{M^{3/2}}{72\gamma^3} + \frac{73}{M^{3/2}}\sigma_h^3.$$

We choose $\gamma \in \arg\min_{\gamma\in(0,\frac{M}{172L_g}]}326(\frac{L_g\gamma}{M})^{9/8}\sigma_g^{3/2} + \frac{M^{3/2}}{72\gamma^3}$, we find $\gamma = \min(\frac{M}{172L_g}, \frac{M^{7/11}}{L_g^{3/11}\sigma_g^{4/11}})$, meaning $\alpha = \min\left(1, 72 \times 32(\frac{L_g^2}{M\sigma_g})^{6/11}\right)$ and get:

$$\frac{1}{1008}\frac{1}{T}\sum_{t=0}^{T-1}\mathbb{E}\mu_M(\boldsymbol{x}_{t+1}) \le \frac{\sqrt{M}F_0}{T} + \frac{M^{6/11}\sigma_g^{6/11}}{192L_g^{12/11}T}\sigma_{g,0}^{3/2} + 327\frac{L_g^{9/11}\sigma_g^{12/11}}{M^{9/22}} + (72 \times 32)^{11/6}\frac{L_g^3}{M^{3/2}} + \frac{73}{M^{3/2}}\sigma_h^3.$$

The term $\frac{L_g^3}{M^{3/2}}$ comes from the constraint $\gamma \le \frac{M}{172L_g}$ being activated which means $\sigma_g^{3/2} \le (72 \times 32)^{11/6}\frac{L_g^3}{M^{3/2}}$. This term limits any possible improvement of the Hessian term (since it has the same power of M), hence why we chose $\beta = 1$ from the start.

Finally, we get

$$\frac{1}{1008}\frac{1}{T}\sum_{t=0}^{T-1}\mathbb{E}\mu_M(\boldsymbol{x}_{t+1}) \le \frac{\sqrt{M}\tilde{F}_{0M}}{T} + \frac{M^{6/11}\sigma_g^{6/11}}{192L_g^{12/11}T}\sigma_{g,0}^{3/2} + 327\frac{L_g^{9/11}\sigma_g^{12/11}}{M^{9/22}} + \frac{\tilde{\sigma}_h^3}{M^{3/2}},$$

where $\tilde{F}_{0M} = F_0 + \frac{M^{1/22}\sigma_g^{6/11}}{192L_g^{12/11}}\sigma_{g,0}^{3/2}$, and $\tilde{\sigma}_h^3 = (72 \times 32)^{11/6}L_g^3 + 73\sigma_h^3$.

□

Note here that the power of M of the gradient term (meaning $\frac{L_g^{9/11}\sigma_g^{12/11}}{M^{9/22}}$) is better than the one we had using the IT momentum in (23) ($\frac{9}{22} - \frac{3}{8} = \frac{6}{176} > 0$), thus we should expect to have a better dependence on T for this term.

By sampling a large first batch (for example $b_g = T/100$ for the gradient such that we can ensure $\sigma_{g,0} \leq \frac{10\sigma_g}{\sqrt{T}}$, and

choosing $M = \max(L, \frac{L_g^{9/10}\sigma_g^{6/5}T^{11/10}}{F_0^{11/10}}, \frac{\tilde{\sigma}_h^2 T^{1/2}}{F_0^{1/2}})$, we get:

$$\frac{1}{1008}\frac{1}{T}\sum_{t=0}^{T-1}\mathbb{E}\mu_M(\boldsymbol{x}_{t+1}) \leq \frac{\sqrt{M}F_0}{T} + 327\frac{L_g^{9/11}\sigma_g^{12/11}}{M^{9/22}} + \frac{\tilde{\sigma}_h^3}{M^{3/2}} + \frac{M^{6/11}\sigma_g^{45/22}}{192L_g^{12/11}T^{7/4}}$$

$$\leq \frac{\sqrt{L}F_0}{T} + 328\left(\frac{L_gF_0}{T}\right)^{9/20}\sigma_g^{6/10} + \left(\frac{F_0\tilde{\sigma}_h}{T}\right)^{3/4}.$$

Note that $\frac{1}{T^{9/20}} \leq \frac{1}{T^{3/7}}$, which means that the MVR momentum slightly improves the gradient rate of stochastic cubic Newton compared to the IT momentum, but this comes at the cost of a slower rate for the Hessian term.