A Variational Perspective on High-Resolution ODEs

Anonymous Author(s)
Affiliation
Address
email

Abstract

We consider unconstrained minimization of smooth convex functions. We propose a novel variational perspective using forced Euler-Lagrange equation that allows for studying high-resolution ODEs. Through this, we obtain a faster convergence rate for gradient norm minimization using Nesterov's accelerated gradient method. Additionally, we show that Nesterov's method can be interpreted as a rate-matching discretization of an appropriately chosen high-resolution ODE. Finally, using the results from the new variational perspective, we propose a stochastic method for noisy gradients. Several numerical experiments compare and illustrate our stochastic algorithm with state of the art methods.

1 Introduction

Smooth convex minimization is a fundamental subclass of optimization problems with broad applications and a rich theoretical foundation that enables the development of powerful optimization methods. In fact, the theory and methods developed for other problem classes, such as non-smooth convex or smooth non-convex optimization, often build upon the work done in smooth convex optimization. As a result, the study of smooth convex minimization has a significant impact on the broader field of optimization.

For the last two decades, first-order methods (*i.e.,* methods that only use gradient information; unlike, for instance, Newton's method that also requires Hessian information) have seen a lot of interest both in theory and applications due to their efficiency and adaptability for large-scale data-driven applications. Gradient descent is one of the oldest and simplest of first-order methods. With a suitable step-size, gradient descent ensures a suboptimality gap (objective residual) of order $\mathcal{O}(1/k)$ after k iterations.

In his seminal work, Nesterov [1983] has shown that the gradient method can achieve faster rates by incorporating momentum deviations. The NAG algorithm ensures a convergence rate of $\mathcal{O}(1/k^2)$ which is an order of magnitude faster than the gradient descent. Remarkably, this rate matches information-theoretical lower bounds for first-order oracle complexity, meaning that NAG is optimal and no other first-order method can guarantee a faster convergence rate [Nesterov, 2003].

The original proof of Nesterov [1983], known as the *estimate sequence technique*, is a highly algebraic and complex procedure difficult to interpret, and provides arguably limited insight into why momentum deviations help with the convergence rates [Hu and Lessard, 2017]. Therefore, many researchers have tried to provide a better understanding of momentum-based acceleration through different perspectives. For example, Su et al. [2016], Shi et al. [2019, 2021], Sanz Serna and Zygalakis [2021] consider a continuous-time perspective; Lessard et al. [2016], Fazlyab et al. [2018] use integral quadratic constraints and control systems with non-linear feedbacks; Muehlebach and Jordan [2019, 2022, 2023] present a dynamical perspective; Attouch et al. [2020, 2021] utilize inertial dynamic involving both viscous damping and Hessian-driven damping; Zhu and Orecchia [2014] views acceleration as a linear coupling of gradient descent and mirror descent updates; and

Submitted to 37th Conference on Neural Information Processing Systems (NeurIPS 2023). Do not distribute.

Ahn and Sra [2022] provides an understanding of the NAG algorithm through an approximation of the proximal point method.

This paper specifically focuses on the *continuous-time perspective*. In [Su et al., 2016], the authors derive a second-order ordinary differential equation (ODE) that exhibits trajectories similar to those of the NAG algorithm in the limit of an infinitesimal step-size $s \to 0$. This result has inspired researchers to analyze various ODEs and discretization schemes to gain a better understanding of the acceleration phenomenon. Notably, Wibisono et al. [2016] demonstrate that the ODE presented in [Su et al., 2016] is a special case of a broader family of ODEs that extend beyond Euclidean space. They achieve this by minimizing the action on a Lagrangian that captures the properties of the problem template, an approach known as the *variational perspective*, and they discretize their general ODE using the *rate-matching* technique. In a related vein, Shi et al. [2021] proposed substituting the low-resolution ODEs (LR-ODEs) introduced in [Su et al., 2016] with high-resolution ODEs (HR-ODEs) which can capture trajectories of NAG more precisely.

Our main contribution in this paper is a novel and innovative extension of the variational perspective for HR-ODEs. A direct combination of these two frameworks is challenging, as it remains unclear how the Lagrangian should be modified to recover HR-ODEs. To address this problem, we propose an alternative approach that preserves the Lagrangian but extends the variational perspective. More specifically, instead of relying on the conventional Euler-Lagrange equation, we leverage the forced Euler-Lagrange equation that incorporates external forces acting on the system. By representing the damped time derivative of the potential function gradients as an external force, our proposed variational perspective allows us to reconstruct various HR-ODEs through specific damping parameters. More details are provided in Section 2.

Other contributions of our paper are as follows: In Section 3, we show that our proposed variational analysis yields a special representation of NAG leading to superior convergence rates than [Shi et al., 2019] in terms of gradient norm minimization. In Section 4, we propose a HR-ODE based on the rate-matching technique. We demonstrate that NAG can be interpreted as an approximation of the rate-matching technique applied to a specific ODE. Furthermore, in Section 5, we extend our analysis to a stochastic setting where gradients are noisy. The proposed stochastic method ensures convergence rates of $\tilde{\mathcal{O}}(1/k^{1/2})$ and $\tilde{\mathcal{O}}(1/k^{3/4})$ for the expected objective residual and the expected gradient norm, respectively. Finally, in Section 6, we present numerical experiments to demonstrate the empirical performance of the proposed method and to validate our theoretical findings.

Problem template and notation. We consider a generic unconstrained smooth convex minimization template:

$$\min_{x \in \mathbb{R}^n} f(x) \tag{1}$$

where $f : \mathbb{R}^n \to \mathbb{R}$ is convex and L-smooth, meaning that its gradient is Lipschitz continuous:

$$\|\nabla f(x) - \nabla f(y)\| \leq L\|x - y\|, \quad \forall x, y \in \mathbb{R}^n, \tag{2}$$

with $\|\cdot\|$ denoting the Euclidean norm. We denote the class of L-smooth convex functions by $\mathcal{F}_L$.

Throughout, we assume that the solution set for Problem (1) is non-empty, and we denote an arbitrary solution by x^*, hence $f^* := f(x^*) \leq f(x)$ for all $x \in \mathbb{R}^n$.

The NAG Algorithm. Given an initial state $x_0 = y_0 \in \mathbb{R}^n$ and a step-size parameter $s > 0$, the NAG algorithm updates the variables x_k and y_k iteratively as follows:

$$\begin{aligned} y_{k+1} &= x_k - s\nabla f(x_k), \\ x_{k+1} &= y_{k+1} + \frac{k}{k+3}(y_{k+1} - y_k). \end{aligned} \tag{NAG}$$

2 External Forces and High-Resolution ODEs

Consider the Lagrangian

$$\mathcal{L}(X_t, \dot{X}_t, t) = e^{\alpha_t + \gamma_t}\left(\frac{1}{2}\|e^{-\alpha_t}\dot{X}_t\|^2 - e^{\beta_t} f(X_t)\right). \tag{3}$$

where $\dot{X}_t \in \mathbb{R}^d$ is the first time-derivative of $X(t)$, and $\alpha_t, \beta_t, \gamma_t : \mathbb{T} \to \mathbb{R}$ are continuously differentiable functions of time that correspond to the weighting of velocity, the potential function f, and the overall damping, respectively. Using variational calculus, we define the action for the curves $\{X_t : t \in \mathbb{R}\}$ as the functional $\mathcal{A}(X) = \int_{\mathbb{R}} \mathcal{L}(X_t, \dot{X}_t, t)dt$. In the absence of external forces, a curve is a stationary point for the problem of minimizing the action $\mathcal{A}(X)$ *if and only if* it satisfies the Euler Lagrange equation $\frac{d}{dt}\{\frac{\partial \mathcal{L}}{\partial \dot{X}_t}(X_t, \dot{X}_t, t)\} = \frac{\partial \mathcal{L}}{\partial X}(X_t, \dot{X}_t, t)$. This was used in [Wibisono et al., 2016, Wilson et al., 2021] to calculate the LR-ODEs for convex and strongly convex functions[1]. Note that the Euler-Lagrange equation as written, does not account for an external force, F (which is non-conservative). In this case, the Euler-Lagrange equation should be modified to the forced Euler-Lagrange Equation

$$\frac{d}{dt}\left\{\frac{\partial \mathcal{L}}{\partial \dot{X}_t}(X_t, \dot{X}_t, t)\right\} - \frac{\partial \mathcal{L}}{\partial X_t}(X_t, \dot{X}_t, t) = F, \tag{4}$$

which itself is the result of integration by parts of Lagrange d'Alembert principle [Campos et al., 2021]. Using the Lagrangian (3) we have

$$\frac{\partial \mathcal{L}}{\partial \dot{X}_t}(X_t, \dot{X}_t, t) = e^{\gamma_t}(e^{-\alpha_t}\dot{X}_t), \quad \frac{\partial \mathcal{L}}{\partial X_t}(X_t, \dot{X}_t, t) = -e^{\gamma_t + \alpha_t + \beta_t}(\nabla f(X_t)). \tag{5}$$

Substituting (5) in (4) gives

$$\ddot{X}_t + (\dot{\gamma}_t - \dot{\alpha}_t)\dot{X}_t + e^{2\alpha_t + \beta_t}\nabla f(X_t) = e^{\alpha_t - \gamma_t}F. \tag{6}$$

In what follows, we will present two different choices of the external force F for convex and one for strongly convex functions.

2.1 Convex Functions

First choice $(F = -\sqrt{s}e^{\gamma_t}\frac{d}{dt}[e^{-\alpha_t}\nabla f(X)])$: In this case, (6) gives

$$\ddot{X}_t + (\dot{\gamma}_t - \dot{\alpha}_t)\dot{X}_t + e^{2\alpha_t + \beta_t}\nabla f = -\sqrt{s}e^{\alpha_t}\frac{d}{dt}\left[e^{-\alpha_t}\nabla f(X_t)\right]. \tag{7}$$

It is possible to show the convergence of X_t to x^* and establish a convergence rate for this as shown in the following theorem (proof in Appendix A.1). The proof of this theorem (and the subsequent theorems in this section) is based on the construction of a suitable Lyapunov function for the corresponding ODE (e.g. see [Siegel, 2019, Shi et al., 2019, Attouch et al., 2020, 2021]). This non-negative function attains zero only at the stationary solution of the corresponding ODE and decreases along the trajectory of the ODE [Khalil, 2002]. For this theorem (and the subsequent theorems), we will define a proper Lyapunov function and prove sufficient decrease of the function f along the corresponding ODE trajectory.

Theorem 2.1. *Under the ideal scaling conditions* $\dot{\beta}_t \leq e^{\alpha_t}, \dot{\gamma}_t = e^{\alpha t}$, X_t *in* (7) *will satisfy*

$$f(X_t) - f(x^*) \leq \mathcal{O}(e^{-\beta_t})$$

for $f \in \mathcal{F}_L$.

Now, choosing parameters as

$$\alpha_t = \log(n(t)), \quad \beta_t = \log(q(t)/n(t)), \quad \dot{\gamma}_t = e^{\alpha_t} = n(t), \tag{8}$$

in (7) gives

$$\begin{cases} \ddot{X}_t + (n(t) - \frac{\dot{n}(t)}{n(t)} + \sqrt{s}\nabla^2 f(X_t))\dot{X}_t + (n(t)q(t) - \sqrt{s}\frac{\dot{n}(t)}{n(t)})\nabla f(X_t) = 0, \\ F = -\sqrt{s}e^{\gamma_t}\frac{d}{dt}[e^{-\alpha_t}\nabla f(X)], \end{cases} \tag{9}$$

which reduces to

$$\ddot{X}_t + \left(\frac{p+1}{t} + \sqrt{s}\nabla^2 f(X_t)\right)\dot{X}_t + \left(Cp^2t^{p-2} + \frac{\sqrt{s}}{t}\right)\nabla f(X_t) = 0, \tag{10}$$

by taking $n(t) = \frac{p}{t}, q(t) = Cpt^{p-1}$.

Remark 2.1.1. For $p = 2, C = 1/4$, equation (10) corresponds to the (H-ODE) in [Laborde and Oberman, 2020].

[1] [Wilson et al., 2021], uses different Lagrangian for strongly convex functions, but the methodology is the same.

Second choice ($F = -\sqrt{s}e^{\gamma_t-\beta_t}\frac{d}{dt}\left[e^{-(\alpha_t-\beta_t)}\nabla f(X_t)\right]$) : In this case, replacing F in (6) gives

$$\ddot{X}_t + (\dot{\gamma}_t - \dot{\alpha}_t)\dot{X}_t + e^{2\alpha_t+\beta_t}\nabla f = -\sqrt{s}e^{\alpha_t-\beta_t}\frac{d}{dt}[e^{-(\alpha_t-\beta_t)}\nabla f(X_t)]. \tag{11}$$

We establish the following convergence result, and the proof can be found in Appendix A.2

Theorem 2.2. *Under the modified ideal scaling conditions* $\dot{\beta}_t \le e^{\alpha_t}, \dot{\gamma}_t = e^{\alpha t}, \ddot{\beta}_t \le e^{\alpha_t}\dot{\beta}_t + 2\dot{\alpha}_t\dot{\beta}_t$, X_t *in (11) will satisfy*

$$f(X_t) - f(x^*) \le \mathcal{O}\left(\frac{1}{e^{\beta_t} + \sqrt{s}e^{-2\alpha_t}\dot{\beta}_t}\right),$$

for $f \in \mathcal{F}_L$.

Taking the same parameters as in (8) gives

$$\begin{cases} \ddot{X}_t + (n(t) - \frac{\dot{n}(t)}{n(t)} + \sqrt{s}\nabla^2 f(X_t))\dot{X}_t + (n(t)q(t) - \sqrt{s}(\frac{\dot{n}(t)}{n(t)} - \frac{\dot{q}(t)n(t)-\dot{n}(t)q(t)}{n(t)q(t)}))\nabla f(X_t) = 0, \\ F = -\sqrt{s}e^{\gamma_t-\beta_t}\frac{d}{dt}\left[e^{-(\alpha_t-\beta_t)}\nabla f(X_t)\right]. \end{cases} \tag{12}$$

which reduces to

$$\ddot{X}_t + \left(\frac{p+1}{t} + \sqrt{s}\nabla^2 f(X_t)\right)\dot{X}_t + \left(Cp^2t^{p-2} + \frac{\sqrt{s}(p+1)}{t}\right)\nabla f(X_t) = 0, \tag{13}$$

for $n(t) = p/t, q(t) = Cpt^{p-1}$.

Remark 2.2.1. Note that setting $C = 1/4, p = 2$ will lead to the ODE

$$\ddot{X}_t + \left(\frac{3}{t} + \sqrt{s}\nabla^2 f(X_t)\right)\dot{X}_t + \left(1 + \frac{3\sqrt{s}}{t}\right)\nabla f(X_t) = 0. \tag{14}$$

This ODE was discretized using the Semi-Implicit Euler (SIE) and the Implicit Euler (IE) discretization schemes in [Shi et al., 2019]. The corresponding optimization algorithms were shown to accelerate. In addition, note that the convergence rate proved in Theorem 2.2 is faster than its counterpart in Theorem 2.1.

2.2 Strongly Convex Functions:

Our analysis is applicable to strongly convex functions as well. Consider the Lagrangian proposed in [Wilson et al., 2021] for strongly convex functions

$$\mathcal{L}(X_t, \dot{X}_t, t) = e^{\alpha_t+\beta_t+\gamma_t}\left(\frac{\mu}{2}\|e^{-\alpha_t}\dot{X}_t\|^2 - f(X_t)\right). \tag{15}$$

Then, the forced Euler-Lagrange equation (4) becomes

$$\ddot{X} + (-\dot{\alpha}_t + \dot{\gamma}_t + \dot{\beta}_t)\dot{X} + \frac{1}{\mu}e^{2\alpha_t}\nabla f(X) = \frac{F}{\mu e^{-\alpha_t+\gamma_t+\beta_t}}. \tag{16}$$

Taking $F = -\sqrt{s}e^{\alpha_t+\gamma_t}\frac{d}{dt}(e^{\beta_t}\nabla f(X_t))$ in (16) gives

$$\ddot{X} + (-\dot{\alpha}_t + \dot{\gamma}_t + \dot{\beta}_t)\dot{X} + \frac{1}{\mu}e^{2\alpha_t}\nabla f(X) = \frac{-\sqrt{s}e^{2\alpha_t-\beta_t}\frac{d}{dt}(e^{\beta_t}\nabla f(X_t))}{\mu}. \tag{17}$$

We can establish the following convergence result for X_t in (17) to the unique minimizer x^*. The proof of this result is deferred to Appendix A.3

Theorem 2.3. *Under the modified ideal scaling conditions* $\alpha_t = \alpha$, $\dot{\beta}_t \le e^{\alpha_t}$, $\dot{\gamma}_t = e^{\alpha_t}$, *and* $\dot{\beta}_t \ge 0$ X_t *in (17) satisfies*

$$f(X_t) - f(x^*) \le \mathcal{O}(e^{-\beta_t}) \tag{18}$$

for μ*-strongly convex function* f.

Remark 2.3.1. Taking $\alpha = \log(\sqrt{\mu})$ and $\gamma_t = \beta_t = \sqrt{\mu}t$ in (17) gives the NAG's corresponding HR-ODE

$$\ddot{X}_t + (2\sqrt{\mu} + \sqrt{s}\nabla^2 f(X_t))\dot{X}_t + (1 + \sqrt{\mu s})\nabla f(X_t) = 0, \tag{19}$$

for μ-strongly convex function f as in [Shi et al., 2021].

3 Gradient Norm Minimization of the NAG

One of the implications of our variational study on HR-ODEs in Section 2 was the ODE (14). Reformulating this ODE gives

$$\begin{cases} \dot{X}_t = n(t)(V_t - X_t) - \sqrt{s}\nabla f(X_t) \\ \dot{V}_t = -q(t)\nabla f(X_t) - \sqrt{s}\frac{\dot{q}(t)n(t) - \dot{n}(t)q(t)}{n^2(t)q(t)}\nabla f(X_t). \end{cases} \tag{20}$$

Applying the SIE on (20) for $X(t) \approx X(t_k), V(t) \approx V(t_k), n(t_k) = p/t_k, q(t_k) = Cpt_k^{p-1}$, $p = 2, t_k = k\sqrt{s}$ and $C = 1/4$ gives

$$\begin{cases} x_{k+1} = x_k + \frac{2}{k}(v_k - x_{k+1}) - s\nabla f(x_k), \\ v_{k+1} = v_k - \frac{1}{2}(ks)\nabla f(x_{k+1}) - s\nabla f(x_{k+1}), \end{cases} \tag{21}$$

which is exactly the NAG algorithm. The interpretation of the NAG method as the SIE discretization of (20) has not been discussed before in the literature (see [Ahn and Sra, 2022] for the four most studied representations). It is precisely this connection with the ODE (20) though that inspires our choice of the Lyapunov function which in turn gives rise to a faster convergence rate. The following theorem formulates this result. The proof is in Appendix A.4 and it is based on the discrete Lyapunov analysis of (21). Similar convergence rate was very recently found by [Chen et al., 2022] through *implicit velocity* perspective on HR-ODEs which uses a different Lyapunov analysis than this work.

Theorem 3.1. *Consider the update (21). Then, if $f \in \mathcal{F}_L$ we have*

$$\min_{0\le i\le k-1} \|\nabla f(x_i)\|^2 \le \frac{12}{k^3 s^2}\|x_0 - x^*\|^2,$$

and

$$f(x_k) - f(x^*) \le \frac{2}{sk(k+2)}\|x_0 - x^*\|^2$$

for $0 \le s \le 1/L$, $v_0 = x_0$, and any $x_0 \in \mathbb{R}^n$.

Remark 3.1.1 (Comparison with state of the art). The rate in Theorem 3.1 is improved compared to the previous rate found in [Shi et al., 2021], which is

$$\min_{0\le i\le k} \|\nabla f(x_i)\|^2 \le \frac{8568}{(k+1)^3 s^2}\|x_0 - x^*\|^2,$$

for $0 < s \le 1/(3L)$ and $k \ge 0$.

4 Rate-Matching Approximates the NAG Algorithm

The ODE (10) when $p = 2, C = 1/4$ is equivalent to

$$\begin{cases} \dot{X}_t = \frac{2}{t}(Z_t - X_t) - \sqrt{s}\nabla f(X_t), \\ \dot{Z}_t = -\frac{t}{2}\nabla f(X_t). \end{cases} \tag{22}$$

which is a perturbation of the LR-ODE

$$\begin{cases} \dot{X}_t = \frac{2}{t}(Z_t - X_t), \\ \dot{Z}_t = -\frac{t}{2}\nabla f(X_t). \end{cases} \tag{23}$$

We now show that when the rate-matching technique in [Wibisono et al., 2016] is applied to (23), the final algorithm reveals similar behaviour as (22). This result is then used to approximately recover the NAG method using rate-matching discretization.

Applying the rate-matching discretization on the ODE (23) gives

$$\begin{cases} x_{k+1} = \frac{2}{k+2}z_k + \frac{k}{k+2}y_k, \\ y_k = x_k - s\nabla f(x_k), \\ z_k = z_{k-1} - \frac{1}{2}sk\nabla f(y_k). \end{cases} \tag{24}$$

which has a convergence rate of $\mathcal{O}(1/(sk^2))$ [Wibisono et al., 2016]. In the following proposition, we study the behaviour of (24) in limit of $s \to 0$. The proof is given in Appendix A.5

Proposition 4.1. *The continuous-time behaviour of (24) is approximately*

$$\ddot{X}_t + \left(\frac{3}{t} + \sqrt{s}\nabla f(X_t)\right)\dot{X}_t + \left(1 + \frac{\sqrt{s}}{t}\right)\nabla f(X_t) = 0, \tag{25}$$

which is the the high-resolution ODE (10).

The ODE (25) is the same as (22). In this sense, rate-matching implicitly perturbs the LR-ODE. The question that naturally arises is that when do we recover the HR-ODE (14) (which corresponds to the NAG algortihm through the SIE discretization) from the rate-matching technique? To answer, we will first perturb the LR-ODE (23) in the second line. Then, the rate-matching discretization is applied. Perturbing (23) gives

$$\begin{cases} \dot{X}_t = \frac{2}{t}(Z_t - X_t), \\ \dot{Z}_t = -\frac{t}{2}\nabla f(X_t) - \sqrt{s}\nabla f(X_t). \end{cases} \tag{26}$$

Discretizing (26) using the rate-matching method with $t_k = k\sqrt{s}$ gives

$$\begin{cases} x_{k+1} = \frac{2}{k+2}z_k + \frac{k}{k+2}y_k, \\ y_k = x_k - s\nabla f(x_k), \\ z_k = z_{k-1} - \frac{s}{2}(k+2)\nabla f(y_k), \end{cases} \tag{27}$$

which is extremely close to the NAG algorithm. Indeed, replacing $\nabla f(y_k)$ with $\nabla f(x_k)$ in the third line of (27) gives exactly the NAG method. Typically, x_k and y_k are very close. This is due to x_k and y_k having a difference of order s. Since in continuous time $X(t_k) \approx Y(t_k)$ (due to $s \to 0$), the HR-ODE of (27) is (14). This means that the corresponding HR-ODE of (27) is

$$\begin{cases} \dot{X}_t = \frac{2}{t}(Z_t - X_t) - \sqrt{s}\nabla f(X_t), \\ \dot{Z}_t = -\frac{t}{2}\nabla f(X_t) - \sqrt{s}\nabla f(X_t). \end{cases} \tag{28}$$

which is the perturbed version of (26) and the HR-ODE associated with the NAG algorithm.

5 Stochastic Extentions

In this section, we propose a stochastic variation of (21). We model noisy gradients by adding i.i.d noise e_k to the gradients. Consider the update

$$\begin{cases} x_{k+1} = x_k + \frac{2s_k}{t_k}(v_k - x_{k+1}) - \frac{\beta s_k}{\sqrt{L}}(\nabla f(x_k) + e_k), \\ v_{k+1} = v_k - \frac{1}{2}(t_k s_k + \frac{2s_k\beta}{\sqrt{L}})(\nabla f(x_{k+1}) + e_{k+1}) \end{cases} \tag{29}$$

with $\beta \geq 2$. This update reduces to (21) when $e_k = 0, s_k = \sqrt{s} = \beta/\sqrt{L}, t_k = k\sqrt{s}$. We will refer to (29) as the Noisy NAG (NNAG) algorithm. The NNAG is interesting due to its capability of dealing with perturbed gradients. This is the case in practical methods e.g. SGD [Bottou, 2010], SAG [Schmidt et al., 2017], SAGA [Defazio et al., 2014], SVRG [Johnson and Zhang, 2013], and etc. The following convergence result holds for the NNAG, and its proof is in Appendix A.7

Theorem 5.1. *Consider* $t_k = \sum_{i=1}^{k} s_k, \beta \geq 2, k_0 \geq (\frac{\beta}{1+\frac{1}{8}(\sum_{i=1}^{k}\frac{1}{i^\alpha})^2})^{1/\alpha}$ *and* $s_k = \frac{c}{k^\alpha}$ *with* $c \leq \frac{1}{\sqrt{L}}$ *and* $\frac{3}{4} \leq \alpha < 1$. *Then, if the NNAG method (29) is used to find* $x^* = \arg\min_x f(x)$ *we have*

$$\mathbb{E}[f(x_k)] - f(x^*) \leq \frac{\mathbb{E}[\varepsilon(k_0)] + \frac{\sigma^2 c^4}{(1-\alpha)^2}\left[k_0^{3-4\alpha} - k^{3-4\alpha}\right] + \frac{\sigma^2 c^3 \beta}{2\sqrt{L}(1-\alpha)(3\alpha-2)}\left[k_0^{2-3\alpha} - k^{2-3\alpha}\right] + \frac{\beta^2 c^2 \sigma^2}{2L(2\alpha-1)}\left[k_0^{1-2\alpha} - k^{1-2\alpha}\right]}{\frac{c^2}{4(1-\alpha)^2}((k^{1-\alpha}-1)^2) + \frac{c\beta}{2\sqrt{L}(1-\alpha)}\left(k^{(1-\alpha)}-1\right)}$$

for $\alpha > 3/4$ *and*

$$\mathbb{E}[f(x_k)] - f(x^*) \leq \frac{\mathbb{E}[\varepsilon(k_0)] + 2\sigma^2 c^4\left[\log(\frac{k}{k_0})\right] + \frac{8\sigma^2 c^3 \beta}{\sqrt{L}}\left[k_0^{-1/4} - k^{-1/4}\right] + \frac{\beta^2 c^2 \sigma^2}{L}\left[k_0^{-1/2} - k^{-1/2}\right]}{4c^2\left((k^{1/4}-1)^2\right) + \frac{2c\beta}{\sqrt{L}}\left(k^{1/4}-1\right)} \tag{30}$$

for $\alpha = 3/4$.

Next, we show that slight modifications to the NNAG method gives rise to another stochastic method with a similar convergence rate as the NNAG algorithm, but more transparent proof (see Appendix A.6). This proof results in a convergence rate for $\mathbb{E}\left[\min_{0\leq i\leq k-1}\|\nabla f(x_i)\|^2\right]$ with a rate of $O(\log(k)/k^{(3/4)})$. It remains a future work to show similar result for the NNAG update.

Theorem 5.2. *Consider the conditions of Theorem 5.1. Then, if the NNAG modifies to*

$$\begin{cases} x_{k+1} = x_k + \frac{2s_k}{t_k}(v_k - x_{k+1}) - \frac{s_k}{\sqrt{L}}(\nabla f(x_k) + e_k), \\ v_{k+1} = v_k - \frac{1}{2}((t_k)s_k)(\nabla f(x_{k+1}) + e_{k+1}) - s_k^2(\nabla f(x_{k+1}) + e_{k+1}). \end{cases} \tag{31}$$

and (31) is used to find $x^* = \arg\min_x f(x)$, we will have

$$\mathbb{E}[f(x_k)] - f(x^*) \leq \begin{cases} \frac{\mathbb{E}[\varepsilon(0)] + \frac{c^4\sigma^2}{8}[16(1+\log(k))+32+6]}{2c^2\left[2(k^{\frac{1}{4}}-1)^2 + k^{-\frac{3}{4}}(k^{\frac{1}{4}}-1)\right]} & \alpha = \frac{3}{4} \\ \frac{\mathbb{E}[\varepsilon(0)] + \frac{c^4\sigma^2}{8}\left[\frac{(4\alpha-2)}{(1-\alpha)^2(4\alpha-3)} + \frac{4(4\alpha-1)}{(1-\alpha)(4\alpha-2)} + \frac{4(4\alpha)}{(4\alpha-1)}\right]}{\frac{c^2}{2(1-\alpha)}\left[\frac{(k^{1-\alpha}-1)^2}{2(1-\alpha)} + k^{-\alpha}(k^{(1-\alpha)}-1)\right]} & 1 > \alpha > \frac{3}{4} \end{cases}, \tag{32}$$

with $\mathbb{E}[\varepsilon(0)] = \frac{1}{2}\|v_0 - x^*\|^2$. In addition, for $\alpha = 3/4$ we have

$$\mathbb{E}\left[\min_{0\leq i\leq k-1}\|\nabla f(x_i)\|^2\right] \leq \frac{2\sqrt{L}\mathbb{E}[\varepsilon(0)] + (2c^4\sigma^2\sqrt{L})(2\log(k) + 6 + \frac{3}{4})}{16c^3\left(\frac{k^{3/4}-1}{3} + k^{1/4} - \frac{3}{2} + k^{1/2}\right)}. \tag{33}$$

Remark 5.2.1. The algorithm (31) reduces to (21) when $e_k = 0, s_k = \sqrt{s} = 1/\sqrt{L}, t_k = k\sqrt{s}$.

Remark 5.2.2 (Connection to the NAG). Note that when the noise is omitted ($\sigma = 0$), the parameter α can become zero. This is because we do not have to alleviate the effect of noise with decreasing step-sizes. Therefore, we recover the convergence rate of $O(1/k^2)$ for the NAG method when $c = 1/\sqrt{L}$.

Remark 5.2.3 (Comparison with [Laborde and Oberman, 2020]). Laborde et al., proposed a stochastic method with noisy gradients. Their method uses another presentation of the NAG algorithm (the presentation from EE discretization). Our rate (32) has the same order of convergence as [Laborde and Oberman, 2020]. However, their analysis did not achieve the bound (33) (see [Laborde and Oberman, 2020] Appendix C.4).

Remark 5.2.4 (Comparison between Theorems 5.1 and 5.2). The rate in (30) is asymptotically similar to (32). However, the transient behaviour of (30) is faster than both (32) and the rate in [Laborde and Oberman, 2020] when L is large (see Figure 1a). This is due to the tuning parameter β which is usually set to L or higher. This scenario (Large L) often happens in practice, e.g. in training a two-layer Convolutional Neural Network (CNN) [Shi et al., 2022].

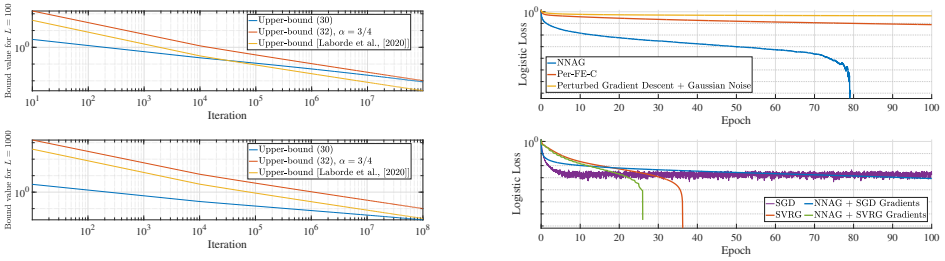
Our convergence result for the NNAG holds after k_0 iterations. Practically, k_0 is lower than $\left(\frac{\beta}{1+\frac{1}{8}(\sum_{i=1}^{k}\frac{1}{i^\alpha})^2}\right)^{1/\alpha}$ and it would be interesting to explore possible lower bounds for k_0.

6 Numerical Results

In this section, we will propose our empirical results. We have separated our simulations into three parts.

Upper Bounds. First, we compare the bounds (30), (32), and Proposition 4.5 in [Laborde and Oberman, 2020]. Figure 1a depicts the result. In practical scenarios where L is large [Shi et al., 2022] the bound (30) is lower than the other two for large enough iterations. This observation has encouraged us to analyse the behaviour of NNAG in practical scenarios e.g. binary classification and CNN training tasks.

Binary Classification. For this task we considered $d = 1000$ randomly generated samples of dimension $n = 10$ and labels. Then, the problem is $\min_x \frac{1}{d}\sum_{i=1}^{d}\log(1 + e^{-y_i\langle x_i, x\rangle})$ where y_i and x_i denote the ith label and sample. For comparison, we considered the perturbed gradient descent method with Gaussian noise and decreasing step-size $s_k = 1/(\sqrt{L}k^{2/3})$ (Proposition 3.4 in [Laborde and Oberman, 2020]) together with accelerated noisy gradient descent (Per-FE-C) in



(a) (b)

Figure 1: (a) Comparison of the upper bounds in this work with state of the art, (b) performance of various methods discussed in this work in binary classification problem.

[Laborde and Oberman, 2020]. For the NNAG we used $s_k = 1/(\sqrt{L}k^{3/4}), \beta = L/10$. All the perturbation was done using i.i.d Gaussian noise with variance equal to one and we conducted 100 Monte-Carlo runs. The result is in Figure 1b top. As shown, NNAG outperforms all the other methods in this case. In a related experiment, we mixed NNAG with SGD and SVRG. For the SGD-mixing we replaced the noisy gradients with SGD-like gradients and for the SVRG-mixing we evaluated all the gradients at the beginning of each epoch (like taking a snapshot in SVRG) and set $t_k = 0$. The step-sizes for SVRG[2], SGD, NNAG+SVRG, and NNAG+SGD were set to $1/(10L)$, $1/L$, $c = 1/\sqrt{L}, \beta = L/10$, and $c = 1/\sqrt{L}, \beta = L$. Also, we conducted 100 Monte-Carlo runs. The result is shown in Figure 1b bottom. When mixed with SGD or SVRG, the NNAG performs better than the original methods. This stems from the generality of NNAG in terms of gradient noise and shows the potential of NNAG to mix with different methods and accelerate them.

Classification on CIFAR10. Here we consider the non-convex task of training a CNN on CIFAR10 dataset [Krizhevsky et al., 2009] using the SGD, SVRG, NNAG, and the NNAG+SVRG methods. The network consisted of two convolutional layers each followed by max pooling and 3 fully connected linear layers each followed by ReLU activation function. The step-sizes for SGD and SVRG were 0.01 and for the NNAG and the NNAG+SVRG algorithms we had $c = 0.05, \beta = 150^2$ and $c = 0.001, \beta = 100^2/10$. The division by 10 is due to step-size division by 10 in the SVRG method. The results for 20 Monte-Carlo simulations are depicted in Figure 2. As the figure depicts, the SVRG+NNAG performs faster than the other methods in terms of minimizing the training error. Interestingly, the validation accuracy of the NNAG is slightly better than the rest which suggests the convergence of the NNAG to another local minima.

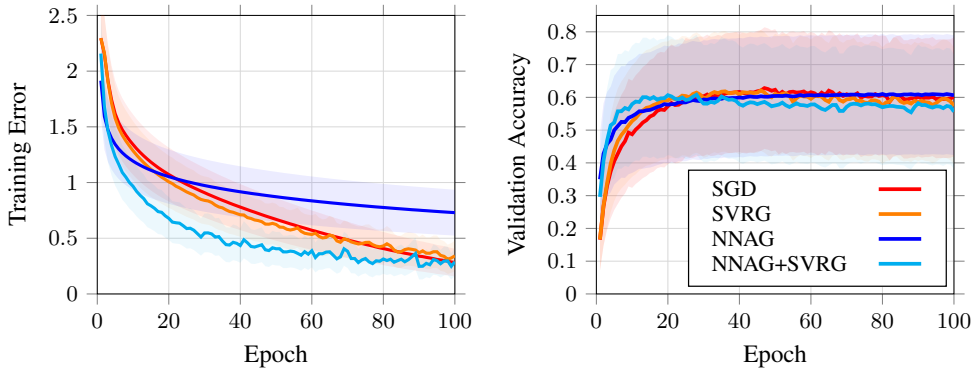


Figure 2: Training error and validation accuracy of NNAG, SGD, SVRG, and NNAG+SVRG when used for training a simple CNN on CIFAR10 dataset. Lower and upper confidence bounds with significance level of 0.68 are drawn with similar color to their corresponding line.

[2]according to the original implementation [Johnson and Zhang, 2013]

7 Related Work

Polyak's Heavy-Ball (HB) method was one of the first momentum-based methods which could accelerate relative to the gradient descent method [Polyak, 1963]. However, this was not the case for every smooth function [Lessard et al., 2016]. Nesterov modified the HB and introduced the NAG method. This method achieved global convergence with a rate of $\mathcal{O}(1/k^2)$ for smooth convex functions [Nesterov, 1983]. Nesterov used estimate sequence technique to show the convergence of the NAG method. This technique does not provide immediate insights toward the success of the NAG algorithm in acceleration. Thus, many have tried different approaches to understand the essence of acceleration.

On similar line of work to continuous-time analysis, [Wibisono et al., 2016] introduced a variational perspective on accelerated methods. This led to a general (non-Euclidean) ODE which contained the ODE found by [Su et al., 2016] as a special case. Their work was based on the choice of a Lagrangian and its corresponding parameters. Since the choice of Lagrangian was not unique, [Wilson et al., 2021] provided a variational perspective on different accelerated first-order methods using a second Lagrangian. [Fazlyab et al., 2017] found a family of accelerated dual algorithms for constrained convex minimization problem through a similar variational approach. Recently, [Zhang et al., 2021] showed that the second-variation also plays an important role in optimality of the ODE found by [Su et al., 2016]. Specifically, they showed that if the time duration is long enough, then the mentioned ODE for the NAG algorithm is the saddle point to the problem of minimizing the action functional.

The dynamical system perspective on NAG was studied in [Muehlebach and Jordan, 2019]. They showed that the NAG is recovered from the SIE discretization of an ODE. The mentioned ODE was not the result of a vanishing step-size argument. They found that a curvature-dependent damping term accounts for the acceleration phenomenon. Interestingly, [Chen et al., 2022] also used similar ODE without the SIE discretization. They showed that implicit-velocity is the reason of the acceleration. In a recent analysis, [Muehlebach and Jordan, 2023] explores the connections between non-smooth dynamical systems and first-order methods for constrained optimization.

8 Conclusion

In this work, we considered unconstrained smooth convex minimization problem in the Euclidean space. Through a variational analysis on HR-ODEs, better convergence rate for the gradient norm minimization of the NAG algorithm was achieved. In addition, we showed that the NAG method can be seen as an approximation of the rate-matching technique when applied on a special ODE. Our analysis was then extended to stochastic scenarios. In particular, we proposed a method with both constant and varying step-sizes which performed comparable and sometimes better than state of the art methods.

This work entails multiple future directions. Nesterov's oracle complexity lower bound on gradient norm minimization is $\mathcal{O}(k^{-4})$ [Nesterov, 2003]. It remains an open question to see if the NAG method can achieve this rate of convergence for gradient norm minimization. In this work, we noticed that the HR-ODEs follow the same external force structure. In the smooth-strongly convex case, Triple Momentum (TM) method is the fastest known globally convergent method [Van Scoy et al., 2018]. However, the HR-ODE associated with the TM method is not shown to achieve the similar convergence rate as the TM method [Sun et al., 2020]. One could use the external force structure proposed here to find a better convergence rate for the HR-ODE associated with the TM algorithm. In addition, our analysis was confined to the Euclidean space. We believe it is possible to explore non-Euclidean forces using a Bregman Lagrangian as in [Wibisono et al., 2016]. Finally, we blended our noisy stochastic scheme with other known stochastic methods (e.g. SGD and SVRG). This technique improved the performance of those methods. As a future work, one can apply the same technique to other practical methods like ADAM, RMSprop, etc, and study the behaviour of the final algorithm.

References

K. Ahn and S. Sra. Understanding nesterov's acceleration via proximal point method. In *Symposium on Simplicity in Algorithms (SOSA)*, pages 117–130. SIAM, 2022.

H. Attouch, Z. Chbani, J. Fadili, and H. Riahi. First-order optimization algorithms via inertial systems with hessian driven damping. *Mathematical Programming*, pages 1–43, 2020.

H. Attouch, Z. Chbani, J. Fadili, and H. Riahi. Convergence of iterates for first-order optimization algorithms with inertia and hessian driven damping. *Optimization*, pages 1–40, 2021.

L. Bottou. Large-scale machine learning with stochastic gradient descent. In *Proceedings of COMPSTAT'2010: 19th International Conference on Computational StatisticsParis France, August 22-27, 2010 Keynote, Invited and Contributed Papers*, pages 177–186. Springer, 2010.

C. M. Campos, A. Mahillo, and D. M. de Diego. A discrete variational derivation of accelerated methods in optimization. *arXiv preprint arXiv:2106.02700*, 2021.

S. Chen, B. Shi, and Y.-x. Yuan. Gradient norm minimization of nesterov acceleration: $o(1/k^3)$. *arXiv preprint arXiv:2209.08862*, 2022.

A. Defazio, F. Bach, and S. Lacoste-Julien. Saga: A fast incremental gradient method with support for non-strongly convex composite objectives. *Advances in neural information processing systems*, 27, 2014.

M. Fazlyab, A. Koppel, V. M. Preciado, and A. Ribeiro. A variational approach to dual methods for constrained convex optimization. In *2017 American Control Conference (ACC)*, pages 5269–5275, 2017. doi: 10.23919/ACC.2017.7963773.

M. Fazlyab, A. Ribeiro, M. Morari, and V. M. Preciado. Analysis of optimization algorithms via integral quadratic constraints: Nonstrongly convex problems. *SIAM Journal on Optimization*, 28(3):2654–2689, 2018. doi: 10.1137/17M1136845. URL https://doi.org/10.1137/17M1136845.

B. Hu and L. Lessard. Dissipativity theory for nesterov's accelerated method. In *International Conference on Machine Learning*, pages 1549–1557. PMLR, 2017.

R. Johnson and T. Zhang. Accelerating stochastic gradient descent using predictive variance reduction. *Advances in neural information processing systems*, 26, 2013.

H. K. Khalil. Nonlinear systems third edition. *Patience Hall*, 115, 2002.

A. Krizhevsky, G. Hinton, et al. *Learning multiple layers of features from tiny images*. Toronto, ON, Canada, 2009.

M. Laborde and A. Oberman. A lyapunov analysis for accelerated gradient methods: from deterministic to stochastic case. In S. Chiappa and R. Calandra, editors, *Proceedings of the Twenty Third International Conference on Artificial Intelligence and Statistics*, volume 108 of *Proceedings of Machine Learning Research*, pages 602–612. PMLR, 26–28 Aug 2020. URL https://proceedings.mlr.press/v108/laborde20a.html

L. Lessard, B. Recht, and A. Packard. Analysis and design of optimization algorithms via integral quadratic constraints. *SIAM J. Optim.*, 26:57–95, 2016.

M. Muehlebach and M. Jordan. A dynamical systems perspective on nesterov acceleration. In *International Conference on Machine Learning*, pages 4656–4662. PMLR, 2019.

M. Muehlebach and M. I. Jordan. On constraints in first-order optimization: A view from non-smooth dynamical systems. *Journal of Machine Learning Research*, 23(256):1–47, 2022.

M. Muehlebach and M. I. Jordan. Accelerated first-order optimization under nonlinear constraints. *arXiv preprint arXiv:2302.00316*, 2023.

Y. Nesterov. A method for solving the convex programming problem with convergence rate $o(1/k^2)$. *Proceedings of the USSR Academy of Sciences*, 269:543–547, 1983.

Y. Nesterov. *Introductory lectures on convex optimization: A basic course*, volume 87. Springer Science & Business Media, 2003.

B. Polyak. Gradient methods for the minimisation of functionals. *Ussr Computational Mathematics and Mathematical Physics*, 3:864–878, 1963.

J. M. Sanz Serna and K. C. Zygalakis. The connections between lyapunov functions for some optimization algorithms and differential equations. *SIAM Journal on Numerical Analysis*, 59(3): 1542–1565, 2021.

M. Schmidt, N. Le Roux, and F. Bach. Minimizing finite sums with the stochastic average gradient. *Mathematical Programming*, 162:83–112, 2017.

B. Shi, S. S. Du, W. J. Su, and M. I. Jordan. Acceleration via symplectic discretization of high-resolution differential equations. *arXiv preprint arXiv:1902.03694*, 2019.

B. Shi, S. S. Du, M. I. Jordan, and W. J. Su. Understanding the acceleration phenomenon via high-resolution differential equations. *ArXiv*, abs/1810.08907, 2021.

Z. Shi, Y. Wang, H. Zhang, J. Z. Kolter, and C.-J. Hsieh. Efficiently computing local lipschitz constants of neural networks via bound propagation. *Advances in Neural Information Processing Systems*, 35:2350–2364, 2022.

J. W. Siegel. Accelerated first-order methods: Differential equations and lyapunov functions. *arXiv preprint arXiv:1903.05671*, 2019.

W. Su, S. Boyd, and E. J. Candès. A differential equation for modeling nesterov's accelerated gradient method: Theory and insights. *Journal of Machine Learning Research*, 17(153):1–43, 2016. URL http://jmlr.org/papers/v17/15-084.html.

B. Sun, J. George, and S. S. Kia. High-resolution modeling of the fastest first-order optimization method for strongly convex functions. In *59th IEEE Conference on Decision and Control, CDC 2020, Jeju Island, South Korea, December 14-18, 2020*, pages 4237–4242. IEEE, 2020. doi: 10.1109/CDC42340.2020.9304444. URL https://doi.org/10.1109/CDC42340.2020.9304444.

B. Van Scoy, R. A. Freeman, and K. M. Lynch. The fastest known globally convergent first-order method for minimizing strongly convex functions. *IEEE Control Systems Letters*, 2(1):49–54, 2018. doi: 10.1109/LCSYS.2017.2722406.

A. Wibisono, A. C. Wilson, and M. I. Jordan. A variational perspective on accelerated methods in optimization. *Proceedings of the National Academy of Sciences*, 113 (47):E7351–E7358, 2016. ISSN 0027-8424. doi: 10.1073/pnas.1614734113. URL https://www.pnas.org/content/113/47/E7351.

A. C. Wilson, B. Recht, and M. I. Jordan. A lyapunov analysis of accelerated methods in optimization. *J. Mach. Learn. Res.*, 22:113–1, 2021.

P. Zhang, A. Orvieto, and H. Daneshmand. Rethinking the variational interpretation of accelerated optimization methods. *Advances in Neural Information Processing Systems*, 34:14396–14406, 2021.

Z. A. Zhu and L. Orecchia. Linear coupling: An ultimate unification of gradient and mirror descent. In *Information Technology Convergence and Services*, 2014.