Unifying First-Order and High-Order Acceleration: A Control-Theoretic Framework

Chris Junchi Li[◇]

Department of Electrical Engineering and Computer Sciences[◇]
University of California, Berkeley

September 2, 2024

Abstract

This paper presents a unified control-theoretic framework for understanding and developing high-order optimization algorithms. By drawing connections between continuous-time and discrete-time dynamical systems, we extend the analysis of first-order methods, such as Nesterov's accelerated gradient algorithm, to the realm of higher-order tensor methods. Our approach introduces a closed-loop control system with inertia and Hessian-driven damping, offering new insights into the acceleration mechanisms in optimization. We establish the existence and uniqueness of solutions for our proposed system and demonstrate that it generalizes well-known results for first-order methods to higher-order settings. Furthermore, we construct a Lyapunov function that facilitates the analysis of convergence rates, revealing that our system achieves optimal rates of convergence in terms of both objective function gap and gradient norm. The implications of our findings are discussed in the context of both theoretical advancements and practical algorithmic design for high-order optimization.

Keywords: High-Order Optimization; Control-Theoretic Approach; Continuous-Time Dynamics; Lyapunov Functions; Convergence Rates

1 Introduction

The study of optimization algorithms has witnessed significant advancements over the past few decades, driven by the need to solve increasingly complex problems in various fields such as machine learning, control theory, and operations research. Central to these developments is the interplay between continuous-time and discrete-time perspectives, which has provided valuable insights into the design and analysis of efficient optimization methods.

Classically, first-order optimization methods like gradient descent and Nesterov's accelerated gradient (NAG) have been widely studied. These methods are often analyzed by interpreting them as discretizations of continuous-time dynamical systems. For instance, the steepest descent and heavy ball methods can be viewed as discretizations of gradient-like dissipative systems. Such interpretations have not only deepened our understanding of these algorithms but have also led to the development of new methods with improved convergence properties.

However, while the continuous-time perspective has been extensively explored for first-order methods, the extension to higher-order optimization algorithms, particularly those involving tensor methods, remains less understood. Existing analyses for high-order methods tend to be more involved and lack the unifying principles that characterize first-order approaches. This gap in understanding motivates the need for a systematic framework that can extend continuous-time analyses to higher-order settings.

In this paper, we introduce a control-theoretic approach to high-order optimization, which unifies first-order and higher-order methods under a common framework. Our approach is based

on the study of a closed-loop control system that incorporates both inertia and Hessian-driven damping. By analyzing this system, we obtain new insights into the acceleration mechanisms for high-order tensor algorithms and establish theoretical guarantees for their convergence rates.

Backgrounds. The interplay between continuous-time and discrete-time perspectives on dynamical systems has made a major impact on optimization theory. Classical examples include (1) the interpretation of steepest descent, heavy ball and proximal algorithms as the explicit and implicit discretization of gradient-like dissipative systems [Polyak(1987), Antipin(1994), Attouch and Cominetti(1996), Alvarez(2000), Attouch et al.(2000), Alvarez and Attouch(2001)]; and (2) the explicit discretization of Newton-like and Levenberg-Marquardt regularized systems [Alvarez and Pérez C(1998), Attouch and Redont(2001), Alvarez et al.(2002), Attouch and Svaiter(2011), Attouch et al.(2012), Maingé(2013), Attouch et al.(2013), Abbas et al.(2014), Attouch et al.(2016a), Attouch and László(2020b), Attouch and László(2020a)], which give standard and regularized Newton algorithms. One particularly salient way that these connections have spurred research is via the use of Lyapunov functions to transfer asymptotic behavior and rates of convergence between continuous time and discrete time.

Recent years have witnessed a flurry of new research focusing on continuous-time perspectives on Nesterov's accelerated gradient algorithm (NAG) [Nesterov(1983)] and related methods [Güler(1992), Beck and Teboulle(2009), Tseng(2010), Nesterov(2013)]. These perspectives arise from derivations that obtain differential equations as limits of discrete dynamics [Su et al.(2016), Krichene et al.(2015), Attouch and Peypouquet(2016), Vassilis et al.(2018), Shi et al.(2018), Muehlebach and Jord Diakonikolas and Orecchia(2019), Attouch and Peypouquet(2019), Sebbouh et al.(2020)], including quasi-gradient formulations and Kurdyka-Lojasiewicz theory [Bégout et al.(2015), Attouch et al.(2020a)] (see the references [Huang(2006), Chergui(2008), Chill and Fašangová(2010), Bárta et al.(2012), Bárta and Fašangová(2016)] for geometrical perspective on the topic), inertial gradient systems with constant or asymptotic vanishing damping [Su et al.(2016), Attouch and Cabot(2017), Attouch et al.(2018), Attouch et al.(2019a)] and their extension to maximally monotone operators [Bot and Csetnek(2016), Attouch and Cabot(2018), Attouch and Cabot(2020)], Hessian-driven damping [Alvarez et al.(2002), Attouch et al.(2012), Attouch et al.(2016b), Shi et al.(2018), Boţ et al.(2020), Attouch et al.(2020b), Attouch et al.(2021a)], time scaling [Attouch et al.(2019a), Attouch et al.(2019c), Attouch et al.(2021a), Attouch et al.(2021b)], dry friction damping [Adly and Attouch(2020), Adly and Attouch(2021)], closed-loop damping [Attouch et al.(2020a), Attouch et al.(2021a)], control-theoretic design [Lessard et al.(2016), Hu and Lessard(2017), Fazlyab et al.(2018)] and Lagrangian and Hamiltonian frameworks [Wibisono et al.(2016), Betancourt et al.(2018), Maddison et al.(2018), O'Donoghue and Maddison(2019), Diakonikolas and Jordan(2020) França et al.(2020), Muehlebach and Jordan(2021), França et al.(2021)]. Examples of hitherto unknown results that have arisen from this line of research include the fact that NAG achieves a fast rate of $o(k^{-2})$ in terms of objective function gap [May(2017), Attouch and Peypouquet(2016), Attouch et al.(2018)] and $O(k^{-3})$ in terms of squared gradient norm [Shi et al.(2018)].

The introduction of the Hessian-driven damping into continuous-time dynamics has been a particular milestone in optimization and mechanics. The precursor of this perspective can be found in the variational characterization of the Levenberg-Marquardt method and Newton's method [Alvarez and Pérez C(19 a development that inspired work on continuous-time Newton-like approaches for convex minimization [Alvarez and Pérez C(1998), Attouch and Redont(2001)] and monotone inclusions [Attouch and Svaiter(2011) Maingé(2013), Attouch et al.(2013), Abbas et al.(2014), Attouch et al.(2016a), Attouch and László(2020b), Attouch and László(2020a)]. Building on these works, [Alvarez et al.(2002)] distinguished Hessian-driven damping from classical continuous Newton formulations and showed its importance in

optimization and mechanics. Subsequently, [Attouch et al.(2016b)] demonstrated the connection between Hessian-driven damping and the forward-backward algorithms in Nesterov acceleration (e.g., FISTA), and combined Hessian-driven damping with asymptotically vanishing damping [Su et al.(2016)]. The resulting dynamics takes the following form:

$$\ddot{x}(t) + \frac{\alpha}{t}\dot{x}(t) + \beta\nabla^2\Phi(x(t))\dot{x}(t) + \nabla\Phi(x(t)) = 0 \tag{1}$$

where it is worth mentioning that the presence of the Hessian does not entail numerical difficulties since it arises in the form $\nabla^2\Phi(x(t))\dot{x}(t)$, which is the time derivative of the function $t \mapsto \nabla\Phi(x(t))$. Further work in this vein appeared in [Shi et al.(2018)], where Nesterov acceleration was interpreted via multiscale limits that yield high-resolution differential equations:

$$\ddot{x}(t) + \frac{3}{t}\dot{x}(t) + \sqrt{s}\nabla^2\Phi(x(t))\dot{x}(t) + \left(1 + \frac{3\sqrt{s}}{2t}\right)\nabla\Phi(x(t)) = 0 \tag{2}$$

These limits were used in particular to distinguish between Polyak's heavy-ball method and NAG, which are not distinguished by naive limiting arguments that yield the same differential equation for both.

Althought the coefficients are different in Eq. (1) and Eq. (2), both contain Hessian-driven damping, which corresponds to a correction term obtained via discretization, and which provides fast convergence to zero of the gradients and reduces the oscillatory aspects. Using this viewpoint, several subtle analyses have been recently provided in work independent of ours [Attouch et al.(2020a), Attouch et al.(2021a)]. In particular, they develop a convergence theory for a general inertial system with asymptotic vanishing damping and Hessian-driven damping. Under certain conditions, the fast convergence is guaranteed in terms of both objective function gap and squared gradient norm. Beyond the aforementioned line of work, however, most of the focus in using continuous-time perspectives to shed light on acceleration has been restricted to the setting of first-order optimization algorithms. As noted in a line of recent work [Monteiro and Svaiter(2013), Nesterov(2018), Arjevani et al.(2019), Gasnikov et al.(2019), Jiang et al.(2019), Bubeck et al.(2019), Song et al.(2021)], there is a significant gap in our understanding of optimal p-th order tensor algorithms with $p \geq 2$, with existing algorithms and analysis being much more involved than NAG.

In this paper, we show that a continuous-time perspective helps to bridge this gap and yields a unified perspective on first-order and higher-order acceleration. We refer to our work as a *control-theoretic perspective*, as it involves the study of a closed-loop control system that can be viewed as a differential equation that is governed by a feedback control law, $\lambda(\cdot)$, satisfying the algebraic equation $(\lambda(t))^p\|\nabla\Phi(x(t))\|^{p-1} = \theta$ for some $\theta \in (0, 1)$. Our approach is similar to that of [Attouch et al.(2013), Attouch et al.(2016a)], for the case without inertia, and it provides a first step into a theory of the autonomous inertial systems that link closed-loop control and optimal high-order tensor algorithms. Mathematically, our system can be written as follows:

$$\ddot{x}(t) + \alpha(t)\dot{x}(t) + \beta(t)\nabla^2\Phi(x(t))\dot{x}(t) + b(t)\nabla\Phi(x(t)) = 0 \tag{3}$$

where (α, β, b) explicitly depends on the variables (x, λ, a), the parameters $c > 0$, $\theta \in (0, 1)$ and the order $p \in \{1, 2, \ldots\}$:

$$\begin{array}{l} \alpha(t) = \frac{2\dot{a}(t)}{a(t)} - \frac{\ddot{a}(t)}{\dot{a}(t)}, \quad \beta(t) = \frac{(\dot{a}(t))^2}{a(t)}, \quad b(t) = \frac{\dot{a}(t)(\dot{a}(t)+\ddot{a}(t))}{a(t)} \\ a(t) = \frac{1}{4}(\int_0^t \sqrt{\lambda(s)}ds + c)^2, \quad (\lambda(t))^p\|\nabla\Phi(x(t))\|^{p-1} = \theta \end{array} \tag{4}$$

The initial condition is $x(0) = x_0 \in \{x \in \mathbb{R}^d \mid \|\nabla\Phi(x)\| \neq 0\}$ and $\dot{x}(0) \in \mathbb{R}^d$. Note that this condition is not restrictive since $\|\nabla\Phi(x_0)\| = 0$ implies that the optimization problem has been already solved. A key ingredient in our system is the algebraic equation $(\lambda(t))^p\|\nabla\Phi(x(t))\|^{p-1} = \theta$, which links the feedback control law $\lambda(\cdot)$ and the gradient norm $\|\nabla\Phi(x(\cdot))\|$, and which generalizes an equation appearing in [Attouch et al.(2016a)] for modeling the proximal Newton algorithm. We recall that Eq. (3) has also been studied in [Attouch et al.(2020a), Attouch et al.(2021a)], who provide a general convergence result when (α, β, b) satisfies certain conditions. However, when $p \geq 2$, the specific choice of (α, β, b) in Eq. (4) does not have an analytic form and it thus seems difficult to verify whether (α, β, b) in our control system satisfies that condition (see [Attouch et al.(2021a), Theorem 2.1])). This topic is beyond the scope of this paper and we leave its investigation to future work.

1.1 Our contribution

Throughout the paper, unless otherwise indicated, we assume that

$$\Phi : \mathbb{R}^d \to \mathbb{R} \textit{ is convex and twice continuously differentiable and the set of global minimizers of } \Phi \text{ is nonempty.}$$

As we shall see, our main results on the existence and uniqueness of solutions and convergence properties of trajectories are valid under this general assumption. We also believe that this general setting paves the way for extensions to nonsmooth convex functions or maximal monotone operators (replacing the gradient by the subdifferential or the operator) [Alvarez et al.(2002), Attouch et al.(2012), Attouch et al.(2016b)]. This is evidenced by the equivalent first-order reformulations of our closed-loop control system in time and space (without the occurrence of the Hessian). However, we do not pursue these extensions in the current paper.

The main contributions of our work are the following:

(i) We study the closed-loop control system of Eq. (3) and Eq. (4) and prove the existence and uniqueness of a local solution. We show that when $p = 1$ and $c = 0$, our feedback law reduces to $\lambda(t) = \theta$ and our overall system reduces to the high-resolution differential equation studied in [Shi et al.(2018)], showing explicitly that our system extends the high-resolution framework from first-order optimization to high-order optimization.

(ii) We construct a simple yet nontrivial Lyapunov function that allows us to establish the existence and uniqueness of a global solution under regularity conditions (see Theorem 1). We also use the Lyapunov function to analyze the convergence rates of the solution trajectories; in particular, we show that the convergence rate is $O(t^{-(3p+1)/2})$ in terms of objective function gap and $O(t^{-3p})$ in terms of squared gradient norm.

(iii) We provide two algorithmic frameworks based on the implicit discretization of our closed-looped control system, one of which generalizes the large-step A-HPE in [Monteiro and Svaiter(2013)]. Our iteration complexity analysis is largely motivated by the aforementioned continuous-time analysis, simplifying the analysis in [Monteiro and Svaiter(2013)] for the case of $p = 2$ and generalizing it to $p > 2$ in a systematic manner (see Theorem 3 and 4 for the details).

(iv) We combine the algorithmic frameworks with an approximate tensor subroutine, yielding a suite of optimal p-th order tensor algorithms for minimizing a convex smooth function Φ which has Lipschitz p-th order derivatives. The resulting algorithms include not only the algorithms

studied in the previous works [Gasnikov et al.(2019), Jiang et al.(2019), Bubeck et al.(2019)] but also yield a new optimal p-th order tensor algorithm. A highlight of our analysis is to show that all these p-th order optimal algorithms minimize the squared gradient norm at a rate of $O(k^{-3p})$, complementing the recent analysis in the aforementioned works.

Organization. The remainder of the paper is organized as follows. In Section 3, we study the closed-loop control system in Eq. (3) and Eq. (4) and prove the existence and uniqueness of a local solution using the Banach fixed-point theorem. In Section 4, we show that our system permits a simple yet nontrivial Lyapunov function which allows us to establish the existence and uniqueness of a global solution and derive convergence rates of solution trajectories. Our iteration complexity analysis is largely motivated by the continuous-time analysis of our system, demonstrating that these algorithms achieve fast gradient minimization. In Section 5, we conclude our work with a brief discussion on future research directions. In the appendix, we provide in Section A two conceptual algorithmic frameworks based on the implicit discretization of our closed-loop control system as well as specific optimal p-th order tensor algorithms, followed by deferred proofs and discussions.

Notation. We use bold lower-case letters such as x to denote vectors, and upper-case letters such as X to denote tensors. For a vector $x \in \mathbb{R}^d$, we let $\|x\|$ denote its ℓ_2 Euclidean norm and let $\mathbb{B}_\delta(x) = \{x' \in \mathbb{R}^d \mid \|x' - x\| \leq \delta\}$ denote its δ-neighborhood. For a tensor $X \in \mathbb{R}^{d_1 \times \cdots \times d_p}$, we define

$$X[z^1, \cdots, z^p] = \sum_{1 \leq i_j \leq d_j, 1 \leq j \leq p} \left[X_{i_1,\cdots,i_p}\right] z_{i_1}^1 \cdots z_{i_p}^p$$

and denote by $\|X\|_{\text{op}} = \max_{\|z^i\|=1, 1 \leq j \leq p} X[z^1, \cdots, z^p]$ its operator norm.

Fix $p \geq 1$, we define $\mathcal{F}_\ell^p(\mathbb{R}^d)$ as the class of convex functions on $\mathbb{R}^d$ with ℓ-Lipschitz p-th order derivatives; that is, $f \in \mathcal{F}_\ell^p(\mathbb{R}^d)$ if and only if f is convex and $\|\nabla^{(p)} f(x') - \nabla^{(p)} f(x)\|_{\text{op}} \leq \ell \|x' - x\|$ for all $x, x' \in \mathbb{R}^d$ in which $\nabla^{(p)} f(x)$ is the p-th order derivative tensor of f at $x \in \mathbb{R}^d$. More specifically, for $\{z^1, z^2, \ldots, z^p\} \subseteq \mathbb{R}^d$, we have

$$\nabla^{(p)} f(x)[z^1, \ldots, z^p] = \sum_{1 \leq i_1, \cdots, i_p \leq d} \left[\frac{\partial^p f}{\partial x_{i_1} \cdots \partial x_{i_p}}(x)\right] z_{i_1}^1 \cdots z_{i_p}^p$$

Given a tolerance $\epsilon \in (0, 1)$, the notation $a = O(b(\epsilon))$ stands for an upper bound, $a \leq C b(\epsilon)$, in which $C > 0$ is independent of ϵ.

2 Further Related Work

In addition to the aforementioned works, we provide a few additional remarks regarding related work on accelerated first-order and high-order algorithms for convex optimization.

A significant body of recent work in convex optimization focuses on understanding the underlying principle behind Nesterov's accelerated first-order algorithm (NAG) [Nesterov(1983), Nesterov(2018)], with a particular focus on the interpretation of Nesterov acceleration as a temporal discretization of a continuous-time dynamical system [Krichene et al.(2015), Su et al.(2016), Attouch and Peypouquet(2016), May(2017), Attouch et al.(2019b), Attouch et al.(2018), Vassilis et al.(2018), Shi et al.(2018), Attouch et al.(2018) Diakonikolas and Orecchia(2019), Muehlebach and Jordan(2019), Attouch and Peypouquet(2019), Attouch et al.(2019a), Sebbouh et al.(2020), Attouch and Cabot(2020), Attouch et al.(2020b), Attouch et al.(2020

Attouch et al.(2021a), Adly and Attouch(2021)]. A line of new first-order algorithms have been obtained from the continuous-time dynamics by various advanced numerical integration strategies [Scieur et al.(2017), Betancourt et al.(2018), Zhang et al.(2018), Maddison et al.(2018), Shi et al.(2019), Wilson et al.(2019)]. In particular, [Scieur et al.(2017)] showed that a basic gradient flow system and multi-step integration scheme yields a class of accelerated first-order optimization algorithms. [Zhang et al.(2018)] applied Runge-Kutta integration to an inertial gradient system without Hessian-driven damping [Wibisono et al.(2016)] and showed that the resulting algorithm is faster than NAG when the objective function is sufficiently smooth and when the order of the integrator is sufficiently large. [Maddison et al.(2018)] and [França et al.(2020)] both considered conformal Hamiltonian systems and showed that the resulting discrete-time algorithm achieves fast convergence under certain smoothness conditions. Very recently, [Shi et al.(2019)] have rigorously justified the use of symplectic Euler integrators compared to explicit and implicit Euler integration, which was further studied by [Muehlebach and Jordan(2021)] and [França et al.(2021)]. Unfortunately, none of these approaches are suitable for interpreting optimal acceleration in high-order tensor algorithms.

Research on acceleration in the second-order setting dates back to Nesterov's accelerated cubic regularized Newton algorithm (ACRN) [Nesterov(2008)] and Monteiro and Svaiter's accelerated Newton proximal extragradient (A-NPE) [Monteiro and Svaiter(2013)]. The ACRN algorithm was extended to a p-th order tensor algorithm with the improved convergence rate of $O(k^{-(p+1)})$ [Baes(2009)] and an adaptive p-th order tensor algorithm with essentially the same rate [Jiang et al.(2020)]. This novel extension was also revisited by [Nesterov(2019)] with a discussion on the efficient implementation of a third-order tensor algorithm. Meanwhile, within the alternative A-NPE framework, a p-th order tensor algorithm was studied in a line of works [Gasnikov et al.(2019), Jiang et al.(2019), Bubeck et al.(2019)] and was shown to achieve a convergence rate of $O(k^{-(3p+1)/2})$, matching the lower bound [Arjevani et al.(2019)]. Subsequently, a high-order coordinate descent algorithm was studied in [Amaral et al.(2020)], and very recently, the high-order A-NPE framework has been specialized to the strongly convex setting [Alves(2021)], generalizing the discrete-time algorithms in this paper with an improved convergence rate. Beyond the setting of Lipschitz continuous derivatives, high-order algorithms and their accelerated variants have been adapted for more general setting with Hölder continuous derivatives [Grapiglia and Nesterov(2017), Grapiglia and Nesterov(2019), Doikov and Nesterov(2019), Grapiglia and Nesterov(2020b), Grapiglia and Nesterov(2020a)] and an optimal algorithm has been proposed in [Song et al.(2021)]. Other settings include structured convex non-smooth minimization [Bullins(2020)], convex-concave minimax optimization and monotone variational inequalities [Bullins and Lai(2020), Ostroukhov et al.(2020)], and structured smooth convex minimization [Nesterov(2020b), Nesterov(2020a), Kamzolov(2020), Kamzolov and Gasnikov(2020)]. In the nonconvex setting, high-order algorithms have been also proposed and analyzed [Birgin et al.(2016), Birgin et al.(2017), Martínez(2017), Cartis et al.(2018), Cartis et al.(2019)].

Unfortunately, the derivations of these algorithms do not flow from a single underlying principle but tend to involve case-specific algebra. As in the case of first-order algorithms, one would hope that a continuous-time perspective would offer unification, but the only work that we are aware of in this regard is [Song et al.(2021)], and the connection to dynamical systems in that work is unclear. In particular, some aspects of the UAF algorithm (see [Song et al.(2021), Algorithm 5.1]), including the conditions in Eq. (5.31) and Eq. (5.32), do not have a continuous-time interpretation but rely on case-specific algebra. Moreover, their continuous-time framework reduces to an inertial system without Hessian-driven damping in the first-order setting, which has been proven to be an inaccurate surrogate as mentioned earlier.

We have been also aware of other type of discrete-time algorithms [Zhang et al.(2018), Maddison et al.(2018), Wilson et al.(2019)] which were derived from continuous-time perspective with theoretical guarantee under certain condition. In particular, [Wilson et al.(2019)] derived a family of first-order algorithms by appeal to the explicit time discretization of the accelerated rescaled gradient dynamics. Their new algorithms are guaranteed to (surprisingly) achieve the same convergence rate as the existing optimal tensor algorithms [Gasnikov et al.(2019), Jiang et al.(2019), Bubeck et al.(2019)]. However, the strong smoothness assumption is necessary and might rule out many interesting application problems. In contrast, all the optimization algorithms developed in this paper are applicable for *general* convex and smooth problems with the optimal rate of convergence.

3 The Closed-Loop Control System

In this section, we study the closed-loop control system in Eq. (3) and Eq. (4). We start by rewriting our system as a first-order system in time and space (without the occurrence of the Hessian) which is important to our subsequent analysis and implicit time discretization. Then, we analyze the algebraic equation $(\lambda(t))^p\|\nabla\Phi(x(t))\|^{p-1} = \theta$ for $\theta \in (0,1)$ and prove the existence and uniqueness of a local solution by appeal to the Banach fixed-point theorem. We conclude by discussing other systems in the literature that exemplify our general framework.

3.1 First-order system in time and space

We rewrite the closed-loop control system in Eq. (3) and Eq. (4) as follows:

$$\ddot{x}(t) + \alpha(t)\dot{x}(t) + \beta(t)\nabla^2\Phi(x(t))\dot{x}(t) + b(t)\nabla\Phi(x(t)) = 0$$

where (α, β, b) explicitly depend on the variables (x, λ, a), the parameters $c > 0$, $\theta \in (0,1)$ and the order $p \in \{1, 2, \ldots\}$:

$$\begin{aligned} &\alpha(t) = \tfrac{2\dot{a}(t)}{a(t)} - \tfrac{\ddot{a}(t)}{\dot{a}(t)}, \quad \beta(t) = \tfrac{(\dot{a}(t))^2}{a(t)}, \quad b(t) = \tfrac{\dot{a}(t)(\dot{a}(t)+\ddot{a}(t))}{a(t)} \\ &a(t) = \tfrac{1}{4}(\textstyle\int_0^t \sqrt{\lambda(s)}ds + c)^2, \quad (\lambda(t))^p\|\nabla\Phi(x(t))\|^{p-1} = \theta \end{aligned}$$

By multiplying both sides of the first equation by $\frac{a(t)}{\dot{a}(t)}$ and using the definition of $\alpha(t)$, $\beta(t)$ and $b(t)$, we have

$$\frac{a(t)}{\dot{a}(t)}\ddot{x}(t) + \left(2 - \frac{a(t)\ddot{a}(t)}{(\dot{a}(t))^2}\right)\dot{x}(t) + \dot{a}(t)\nabla^2\Phi(x(t))\dot{x}(t) + (\dot{a}(t) + \ddot{a}(t))\nabla\Phi(x(t)) = 0$$

Defining $z_1(t) = \frac{a(t)}{\dot{a}(t)}\dot{x}(t)$ and $z_2(t) = \dot{a}(t)\nabla\Phi(x(t))$, we have

$$\dot{z}_1(t) = \frac{a(t)}{\dot{a}(t)}\ddot{x}(t) + \left(1 - \frac{a(t)\ddot{a}(t)}{(\dot{a}(t))^2}\right)\dot{x}(t), \quad \dot{z}_2(t) = \dot{a}(t)\nabla^2\Phi(x(t))\dot{x}(t) + \ddot{a}(t)\nabla\Phi(x(t))$$

Putting these pieces together yields

$$\dot{z}_1(t) + \dot{x}(t) + \dot{z}_2(t) = -\dot{a}(t)\nabla\Phi(x(t))$$

Integrating this equation over the interval $[0, t]$, we have

$$z_1(t) + x(t) + z_2(t) = z_1(0) + x(0) + z_2(0) - \int_0^t \dot{a}(s)\nabla\Phi(x(s))ds \tag{5}$$

Since $x(0) = x_0 \in \{x \in \mathbb{R}^d \mid \|\nabla\Phi(x)\| \neq 0\}$, it is easy to verify that $\lambda(0)$ is well defined and determined by the algebraic equation $\lambda(0) = \theta^{\frac{1}{p}}\|\nabla\Phi(x_0)\|^{-\frac{p-1}{p}}$. Using the definition of $a(t)$, we have $a(0) = \frac{c^2}{4}$ and $\dot{a}(0) = \frac{c\theta^{\frac{1}{2p}}\|\nabla\Phi(x_0)\|^{-\frac{p-1}{2p}}}{2}$. Putting these pieces together with the definition of $z_1(t)$ and $z_2(t)$, we have

$$\begin{aligned} z_1(0) + x(0) + z_2(0) &= \frac{a(0)}{\dot{a}(0)}\dot{x}(0) + x(0) + \dot{a}(0)\nabla\Phi(x(0)) \\ &= x(0) + \frac{c\theta^{-\frac{1}{2p}}\dot{x}(0)\|\nabla\Phi(x(0))\|^{\frac{p-1}{2p}} + c\theta^{\frac{1}{2p}}\|\nabla\Phi(x(0))\|^{-\frac{p-1}{2p}}\nabla\Phi(x(0))}{2} \end{aligned}$$

This implies that $z_1(0) + x(0) + z_2(0)$ is completely determined by the initial condition and parameters $c > 0$ and $\theta \in (0, 1)$. For simplicity, we define $v_0 := z_1(0) + x(0) + z_2(0)$ and rewrite Eq. (5) in the following form:

$$\frac{a(t)}{\dot{a}(t)}\dot{x}(t) + x(t) + \dot{a}(t)\nabla\Phi(x(t)) = v_0 - \int_0^t \dot{a}(s)\nabla\Phi(x(s))ds \tag{6}$$

By introducing a new variable $v(t) = v_0 - \int_0^t \dot{a}(s)\nabla\Phi(x(s))ds$, we rewrite Eq. (6) in the following equivalent form:

$$\dot{v}(t) + \dot{a}(t)\nabla\Phi(x(t)) = 0, \quad \dot{x}(t) + \frac{\dot{a}(t)}{a(t)}(x(t) - v(t)) + \frac{(\dot{a}(t))^2}{a(t)}\nabla\Phi(x(t)) = 0$$

Summarizing, the closed-loop control system in Eq. (3) and Eq. (4) can be written as a first-order system in time and space as follows:

$$\begin{cases} \dot{v}(t) + \dot{a}(t)\nabla\Phi(x(t)) = 0 \\ \dot{x}(t) + \frac{\dot{a}(t)}{a(t)}(x(t) - v(t)) + \frac{(\dot{a}(t))^2}{a(t)}\nabla\Phi(x(t)) = 0 \\ a(t) = \frac{1}{4}(\int_0^t \sqrt{\lambda(s)}ds + c)^2 \\ (\lambda(t))^p\|\nabla\Phi(x(t))\|^{p-1} = \theta \\ (x(0), v(0)) = (x_0, v_0) \end{cases} \tag{7}$$

We also provide another first-order system in time and space with different variable (x, v, λ, γ). We study this system because its implicit time discretization leads to a new algorithmic framework which does not appear in the literature. This first-order system is summarized as follows:

$$\begin{cases} \dot{v}(t) - \frac{\dot{\gamma}(t)}{\gamma^2(t)}\nabla\Phi(x(t)) = 0 \\ \dot{x}(t) - \frac{\dot{\gamma}(t)}{\gamma(t)}(x(t) - v(t)) + \frac{(\dot{\gamma}(t))^2}{(\gamma(t))^3}\nabla\Phi(x(t)) = 0 \\ \gamma(t) = 4(\int_0^t \sqrt{\lambda(s)}ds + c)^{-2} \\ (\lambda(t))^p\|\nabla\Phi(x(t))\|^{p-1} = \theta \\ (x(0), v(0)) = (x_0, v_0) \end{cases} \tag{8}$$

Remark. The first-order systems in Eq. (7) and Eq. (8) are equivalent. It suffices to show that

$$\dot{a}(t) = -\frac{\dot{\gamma}(t)}{\gamma^2(t)}, \quad \frac{\dot{a}(t)}{a(t)} = -\frac{\dot{\gamma}(t)}{\gamma(t)}, \quad \frac{(\dot{a}(t))^2}{a(t)} = \frac{(\dot{\gamma}(t))^2}{(\gamma(t))^3}$$

By the definition of $a(t)$ and $\gamma(t)$, we have $a(t) = \frac{1}{\gamma(t)}$ which implies that $\dot{a}(t) = -\frac{\dot{\gamma}(t)}{\gamma^2(t)}$.

Remark. The first-order systems in Eq. (7) and Eq. (8) pave the way for extensions to nonsmooth convex functions or maximal monotone operators (replacing the gradient by the subdifferential or the operator), as done in [Alvarez et al.(2002)] and [Attouch et al.(2012), Attouch et al.(2016b)]. In this setting, either the open-loop case or the closed-loop case without inertia has been studied in the literature [Attouch and Svaiter(2011), Maingé(2013), Attouch et al.(2013), Abbas et al.(2014), Attouch et al.(2016a), Bot and Csetnek(2016), Attouch and Cabot(2018), Attouch and Cabot(2020), Attouch and László(2020b)], but there is significantly less work on the case of a closed-loop control system with inertia. For recent progress in this direction, see [Attouch et al.(2020a)] and references therein.

3.2 Algebraic equation

We study the algebraic equation,

$$(\lambda(t))^p \|\nabla\Phi(x(t))\|^{p-1} = \theta \in (0,1) \tag{9}$$

which links the feedback control $\lambda(\cdot)$ and the solution trajectory $x(\cdot)$ in the closed-loop control system. To streamline the presentation, we define a function $\varphi : [0,+\infty)\times\mathbb{R}^d \mapsto [0,+\infty)$ such that

$$\varphi(\lambda, x) = \lambda\|\nabla\Phi(x)\|^{\frac{p-1}{p}}, \quad \varphi(0,x) = 0$$

By definition, Eq. (9) is equivalent to $\varphi(\lambda(t), x(t)) = \theta^{1/p}$. Our first proposition presents a property of the mapping $\varphi(\cdot, x)$, for a fixed $x \in \mathbb{R}^d$ satisfying $\nabla\Phi(x) \neq 0$. We have:

Proposition 1. *Fixing $x \in \mathbb{R}^d$ with $\nabla\Phi(x) \neq 0$, the mapping $\varphi(\cdot, x)$ satisfies*

(i) $\varphi(\cdot, x)$ is linear, strictly increasing and $\varphi(0,x) = 0$.

(ii) $\varphi(\lambda, x) \to +\infty$ as $\lambda \to +\infty$.

In view of Proposition 1, for any fixed point x with $\nabla\Phi(x) \neq 0$, there exists a unique $\lambda > 0$ such that $\varphi(\lambda, x) = \theta^{1/p}$ for some $\theta \in (0,1)$. We accordingly define $\Omega \subseteq \mathbb{R}^d$ and the mapping $\Lambda_\theta : \Omega \mapsto (0,\infty)$ as follows:

$$\Omega = \{x \in \mathbb{R}^d \mid \|\nabla\Phi(x)\| \neq 0\}, \quad \Lambda_\theta(x) = \theta^{\frac{1}{p}}\|\nabla\Phi(x)\|^{-\frac{p-1}{p}} \tag{10}$$

We now provide several basic results concerning Ω and $\Lambda_\theta(\cdot)$ which are crucial to the proof of existence and uniqueness presented in the next subsection.

Proposition 2. *The set Ω is open.*

Proposition 3. *Fixing $\theta \in (0,1)$, the mappings $\Lambda_\theta(\cdot)$ and $\sqrt{\Lambda_\theta(\cdot)}$ are continuous and locally Lipschitz over Ω.*

3.3 Existence and uniqueness of a local solution

We prove the existence and uniqueness of a local solution of the closed-loop control system in Eq. (3) and Eq. (4) by appeal to the Banach fixed-point theorem. Using the results in Section 3.1 (see Eq. (6)), our system can be equivalently written as follows:

$$\begin{cases} \dot{x}(t) + \frac{\dot{a}(t)}{a(t)}(x(t) + \int_0^t \dot{a}(s)\nabla\Phi(x(s))ds - v_0) + \frac{(\dot{a}(t))^2}{a(t)}\nabla\Phi(x(t)) = 0 \\ a(t) = \frac{1}{4}(\int_0^t \sqrt{\lambda(s)}ds + c)^2 \\ (\lambda(t))^p\|\nabla\Phi(x(t))\|^{p-1} = \theta \\ x(0) = x_0 \end{cases}$$

Using the mapping $\Lambda_\theta : \Omega \mapsto (0, \infty)$ (see Eq. (10)), this system can be further formulated as an autonomous system. Indeed, we have

$$\lambda(t) = \Lambda_\theta(x(t)) \Longleftrightarrow \lambda(t)]^p \|\nabla\Phi(x(t))\|^{p-1} = \theta$$

which implies that

$$a(t) = \frac{1}{4}\left(\int_0^t \sqrt{\Lambda_\theta(x(s))}ds + c\right)^2, \quad \dot{a}(t) = \frac{1}{2}\sqrt{\Lambda_\theta(x(t))}\left(\int_0^t \sqrt{\Lambda_\theta(x(s))}\ ds + c\right)$$

Putting these pieces together, we arrive at an autonomous system in the following compact form:

$$\dot{x}(t) = F(t, x(t)), \quad x(0) = x_0 \in \Omega \tag{11}$$

where the vector field $F : [0, +\infty) \times \Omega \mapsto \mathbb{R}^d$ is given by

$$\begin{aligned} F(t, x(t)) \quad = \quad & -\frac{\sqrt{\Lambda_\theta(x(t))}(2x(t)+\int_0^t \sqrt{\Lambda_\theta(x(s))}(\int_0^s \sqrt{\Lambda_\theta(x(w))}dw+c)\nabla\Phi(x(s))ds-v_0)}{\int_0^t \sqrt{\Lambda_\theta(x(s))}\ ds+c} \\ & -\Lambda_\theta(x(t))\nabla\Phi(x(t)) \end{aligned} \tag{12}$$

A common method for proving the existence and uniqueness of a local solution is via appeal to the Cauchy-Lipschitz theorem [Coddington and Levinson(1955), Theorem I.3.1]. This theorem, however, requires that $F(t, x)$ be continuous in t and Lipschitz in x, and this is not immediate in our case due to the appearance of $\int_0^t \sqrt{\Lambda_\theta(x(s))}ds$. We instead recall that the proof of the Cauchy-Lipschitz theorem is generally based on the Banach fixed-point theorem [Granas and Dugundji(2013)], and we avail ourselves directly of the latter theorem. In particular, we construct Picard iterates ψ_k whose limit is a fixed point of a contraction T. We have the following proposition.

Proposition 4. *There exists $t_0 > 0$ such that the autonomous system in Eq.* (11) *and Eq.* (12) *has a unique solution $x : [0, t_0] \mapsto \mathbb{R}^d$.*

4 Lyapunov Function

In this section, we construct a Lyapunov function that allows us to prove existence and uniqueness of a global solution of our closed-loop control system and to analyze convergence rates. As we will see, an analysis of the rate of decrease of the Lyapunov function together with the algebraic equation permit the derivation of new convergence rates for both the objective function gap and the squared gradient norm.

4.1 Existence and uniqueness of a global solution

Our main theorem on the existence and uniqueness of a global solution is summarized as follows.

Theorem 1. *Suppose that λ is absolutely continuous on any finite bounded interval. Then the closed-loop control system in Eq.* (3) *and Eq.* (4) *has a unique global solution, $(x, \lambda, a) : [0, +\infty) \mapsto \mathbb{R}^d \times (0, +\infty) \times (0, +\infty)$.*

Remark. Intuitively, the feedback law $\lambda(\cdot)$, which we will show satisfies $\lambda(t) \to +\infty$ as $t \to +\infty$, links to the gradient norm $\|\nabla\Phi(x(\cdot))\|$ via the algebraic equation. Since we are interested in the worst-case convergence rate of solution trajectories, which corresponds to the worst-case iteration

complexity of discrete-time algorithms, it is necessary that λ does not dramatically change. In open-loop Levenberg-Marquardt systems, [Attouch and Svaiter(2011)] impose the same condition on the regularization parameters. In closed-loop control systems, however, λ is not a given datum but an emergent component of the dynamics. Thus, it is preferable to prove that λ satisfies this condition rather than assuming it, as done in [Attouch et al.(2013), Theorem 5.2] and [Attouch et al.(2016a), Theorem 2.4] for a closed-loop control system without inertia. The key step in their proof is to show that $\lambda(t) \leq \lambda(0)e^t$ locally by exploiting the specific structure of their system. This technical approach is, however, not applicable to our system due to the incorporation of the inertia term; see Section E for further discussion.

Recall that the system in Eq. (3) and Eq. (4) can be equivalently written as the first-order system in time and space, as in Eq. (7). Accordingly, we define the following simple Lyapunov function:

$$\mathcal{E}(t) = a(t)(\Phi(x(t)) - \Phi(x^\star)) + \frac{1}{2}\|v(t) - x^\star\|^2 \tag{13}$$

where $x^\star$ is a global optimal solution of Φ.

Remark. Note that the Lyapunov function (13) is composed of a sum of the mixed energy $\frac{1}{2}\|v(t) - x^*\|$ and the potential energy $a(t)(\Phi(x(t)) - \Phi(x^*))$. This function is similar to Lyapunov functions developed for analyzing the convergence of Newton-like dynamics [Attouch and Svaiter(2011), Attouch et al.(2013), Abbas et al.(2014), Attouch et al.(2016a)] and the inertial gradient system with asymptotic vanishing damping [Su et al.(2016), Attouch et al.(2016b), Shi et al.(2018), Wilson et al.(2021)]. Indeed, [Wilson et al.(2021)] construct a unified time-dependent Lyapunov function using the Bregman divergence and showed that their approach is equivalent to Nesterov's estimate sequence technique in a number of cases, including quasi-monotone subgradient, accelerated gradient descent and conditional gradient. Our Lyapunov function differs from existing choices in that v is not a standard momentum term depending on $\dot{x}$, but depends on x, λ and $\nabla\Phi$; see Eq. (7).

4.2 Rate of convergence

We establish a convergence rate for a global solution of the closed-loop control system in Eq. (3) and Eq. (4).

Theorem 2. *Suppose that* $(x, \lambda, a) : [0, +\infty) \mapsto \mathbb{R}^d \times (0, +\infty) \times (0, +\infty)$ *is a global solution of the closed-loop control system in Eq.* (3) *and Eq.* (4)*. Then, the objective function gap satisfies*

$$\Phi(x(t)) - \Phi(x^\star) = O(t^{-\frac{3p+1}{2}})$$

and the squared gradient norm satisfies

$$\inf_{0\leq s\leq t} \|\nabla\Phi(x(s))\|^2 = O(t^{-3p})$$

Remark. This theorem shows that the convergence rate is $O(t^{-(3p+1)/2})$ in terms of objective function gap and $O(t^{-3p})$ in terms of squared gradient norm. Note that the former result does not imply the latter result but only gives a rate of $O(t^{-(3p+1)/2})$ for the squared gradient norm minimization even when $\Phi \in \mathcal{F}^1_\ell(\mathbb{R}^d)$ is assumed with $\|\nabla\Phi(x(t))\|^2 \leq 2\ell(\Phi(x(t)) - \Phi(x^\star))$. In fact, the squared gradient norm minimization is generally of independent interest [Nesterov(2012), Shi et al.(2018), Grapiglia and Nesterov(2020a)] and its analysis involves different techniques.

5 Conclusions

We have introduced a control-theoretic framework for modeling and analyzing optimal high-order tensor algorithms for smooth convex optimization. Our approach leverages a closed-loop control system governed by the gradient and Hessian of the objective function, coupled with a feedback control law. We established the existence and uniqueness of both local and global solutions to this system, and provided convergence guarantees through the construction of Lyapunov functions. Our framework offers a systematic method for deriving discrete-time p-th order optimal tensor algorithms, demonstrating that these algorithms minimize the squared gradient norm at a rate of $O(k^{-3p})$.

Furthermore, our analysis underscores the significance of the algebraic equation that arises in the $p \geq 2$ setting, which plays a crucial role in achieving optimal acceleration. This work opens several avenues for future research, including the exploration of nonlinear damping, connections with Lagrangian and Hamiltonian frameworks, and extensions to nonsmooth optimization through differential inclusions and monotone operators. Investigating the continuous-time limits of Nesterov's algorithms may also provide further insights into the geometric and dynamical roles of the algebraic equation in optimization.

References

[Abbas et al.(2014)] B. Abbas, H. Attouch, and B. F. Svaiter. Newton-like dynamics and forward-backward methods for structured monotone inclusions in Hilbert spaces. *Journal of Optimization Theory and Applications*, 161(2):331–360, 2014.

[Adly and Attouch(2020)] S. Adly and H. Attouch. Finite convergence of proximal-gradient inertial algorithms combining dry friction with hessian-driven damping. *SIAM Journal on Optimization*, 30(3): 2134–2162, 2020.

[Adly and Attouch(2021)] S. Adly and H. Attouch. First-order inertial algorithms involving dry friction damping. *Mathematical Programming*, pages 1–41, 2021.

[Alvarez(2000)] F. Alvarez. On the minimizing property of a second order dissipative system in Hilbert spaces. *SIAM Journal on Control and Optimization*, 38(4):1102–1119, 2000.

[Alvarez and Attouch(2001)] F. Alvarez and H. Attouch. An inertial proximal method for maximal monotone operators via discretization of a nonlinear oscillator with damping. *Set-Valued Analysis*, 9(1):3–11, 2001.

[Alvarez and Pérez C(1998)] F. Alvarez and J. M. Pérez C. A dynamical system associated with Newton's method for parametric approximations of convex minimization problems. *Applied Mathematics and Optimization*, 38:193–217, 1998.

[Alvarez et al.(2002)] F. Alvarez, H. Attouch, J. Bolte, and P. Redont. A second-order gradient-like dissipative dynamical system with Hessian-driven damping: Application to optimization and mechanics. *Journal de mathématiques pures et appliquées*, 81(8):747–779, 2002.

[Alves(2021)] M. M. Alves. Variants of the A-HPE and large-step a-hpe algorithms for strongly convex problems with applications to accelerated high-order tensor methods. *ArXiv Preprint: 2102.02045*, 2021.

[Amaral et al.(2020)] V. S. Amaral, R. Andreani, E. G. Birgin, D. S. Marcondes, and J. M. Martínez. On complexity and convergence of high-order coordinate descent algorithms. *ArXiv Preprint: 2009.01811*, 2020.

[Antipin(1994)] A. S. Antipin. Minimization of convex functions on convex sets by means of differential equations. *Differential Equations*, 30(9):1365–1375, 1994.

[Arjevani et al.(2019)] Y. Arjevani, O. Shamir, and R. Shiff. Oracle complexity of second-order methods for smooth convex optimization. *Mathematical Programming*, 178(1):327–360, 2019.

[Attouch and Cabot(2017)] H. Attouch and A. Cabot. Asymptotic stabilization of inertial gradient dynamics with time-dependent viscosity. *Journal of Differential Equations*, 263(9):5412–5458, 2017.

[Attouch and Cabot(2018)] H. Attouch and A. Cabot. Convergence of damped inertial dynamics governed by regularized maximally monotone operators. *Journal of Differential Equations*, 264(12):7138–7182, 2018.

[Attouch and Cabot(2020)] H. Attouch and A. Cabot. Convergence of a relaxed inertial proximal algorithm for maximally monotone operators. *Mathematical Programming*, 184(1):243–287, 2020.

[Attouch and Cominetti(1996)] H. Attouch and R. Cominetti. A dynamical approach to convex minimization coupling approximation with the steepest descent method. *Journal of Differential Equations*, 128(2): 519–540, 1996.

[Attouch and László(2020a)] H. Attouch and S. C. László. Continuous Newton-like inertial dynamics for monotone inclusions. *Set-Valued and Variational Analysis*, pages 1–27, 2020a.

[Attouch and László(2020b)] H. Attouch and S. C. László. Newton-like inertial dynamics and proximal algorithms governed by maximally monotone operators. *SIAM Journal on Optimization*, 30(4):3252–3283, 2020b.

[Attouch and Peypouquet(2016)] H. Attouch and J. Peypouquet. The rate of convergence of Nesterov's accelerated forward-backward method is actually faster than 1/kˆ2. *SIAM Journal on Optimization*, 26(3):1824–1834, 2016.

[Attouch and Peypouquet(2019)] H. Attouch and J. Peypouquet. Convergence rate of proximal inertial algorithms associated with Moreau envelopes of convex functions. In *Splitting Algorithms, Modern Operator Theory, and Applications*, pages 1–44. Springer, 2019.

[Attouch and Redont(2001)] H. Attouch and P. Redont. The second-order in time continuous Newton method. In *Approximation, optimization and mathematical economics*, pages 25–36. Springer, 2001.

[Attouch and Svaiter(2011)] H. Attouch and B. F. Svaiter. A continuous dynamical Newton-like approach to solving monotone inclusions. *SIAM Journal on Control and Optimization*, 49(2):574–598, 2011.

[Attouch et al.(2000)] H. Attouch, X. Goudou, and P. Redont. The heavy ball with friction method, I. the continuous dynamical system: global exploration of the local minima of a real-valued function by asymptotic analysis of a dissipative dynamical system. *Communications in Contemporary Mathematics*, 2(01):1–34, 2000.

[Attouch et al.(2012)] H. Attouch, P-E. Maingé, and P. Redont. A second-order differential system with Hessian-driven damping: Application to non-elastic shock laws. *Differential Equations & Applications*, 4(1):27–65, 2012.

[Attouch et al.(2013)] H. Attouch, P. Redont, and B. F. Svaiter. Global convergence of a closed-loop regularized Newton method for solving monotone inclusions in Hilbert spaces. *Journal of Optimization Theory and Applications*, 157(3):624–650, 2013.

[Attouch et al.(2016a)] H. Attouch, M. M. Alves, and B. F. Svaiter. A dynamic approach to a proximal-Newton method for monotone inclusions in Hilbert spaces, with complexity o (1/nˆ 2). *Journal of Convex Analysis*, 23(1):139–180, 2016a.

[Attouch et al.(2016b)] H. Attouch, J. Peypouquet, and P. Redont. Fast convex optimization via inertial dynamics with Hessian driven damping. *Journal of Differential Equations*, 261(10):5734–5783, 2016b.

[Attouch et al.(2018)] H. Attouch, Z. Chbani, J. Peypouquet, and P. Redont. Fast convergence of inertial dynamics and algorithms with asymptotic vanishing viscosity. *Mathematical Programming*, 168(1-2): 123–175, 2018.

[Attouch et al.(2019a)] H. Attouch, Z. Chbani, and H. Riahi. Fast convex optimization via time scaling of damped inertial gradient dynamics. *Pure and Applied Functional Analysis*, To appear, 2019a.

[Attouch et al.(2019b)] H. Attouch, Z. Chbani, and H. Riahi. Rate of convergence of the nesterov accelerated gradient method in the subcritical case $\alpha \leq 3$. *ESAIM: Control, Optimisation and Calculus of Variations*, 25:2, 2019b.

[Attouch et al.(2019c)] H. Attouch, Z. Chbani, and H. Riahi. Fast proximal methods via time scaling of damped inertial dynamics. *SIAM Journal on Optimization*, 29(3):2227–2256, 2019c.

[Attouch et al.(2020a)] H. Attouch, R. I. Bot, and E. R. Csetnek. Fast optimization via inertial dynamics with closed-loop damping. *ArXiv Preprint: 2008.02261*, 2020a.

[Attouch et al.(2020b)] 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, 2020b.

[Attouch et al.(2021a)] H. Attouch, A. Balhag, Z. Chbani, and H. Riahi. Fast convex optimization via inertial dynamics combining viscous and Hessian-driven damping with time rescaling. *Evolution Equations & Control Theory*, To appear, 2021a.

[Attouch et al.(2021b)] H. Attouch, Z. Chbani, J. Fadili, and H. Riahi. Fast convergence of dynamical ADMM via time scaling of damped inertial dynamics. *ArXiv Preprint: 2103.12675*, 2021b.

[Baes(2009)] M. Baes. Estimate sequence methods: extensions and approximations. *Institute for Operations Research, ETH, Zürich, Switzerland*, 2009.

[Bárta and Fašangová(2016)] T. Bárta and E. Fašangová. Convergence to equilibrium for solutions of an abstract wave equation with general damping function. *Journal of Differential Equations*, 260(3): 2259–2274, 2016.

[Bárta et al.(2012)] T. Bárta, R. Chill, and E. Fašangová. Every ordinary differential equation with a strict Lyapunov function is a gradient system. *Monatshefte für Mathematik*, 166(1):57–72, 2012.

[Beck and Teboulle(2009)] A. Beck and M. Teboulle. A fast iterative shrinkage-thresholding algorithm for linear inverse problems. *SIAM Journal on Imaging Sciences*, 2(1):183–202, 2009.

[Bégout et al.(2015)] P. Bégout, J. Bolte, and M. A. Jendoubi. On damped second-order gradient systems. *Journal of Differential Equations*, 259(7):3115–3143, 2015.

[Betancourt et al.(2018)] M. Betancourt, M. I. Jordan, and A. C. Wilson. On symplectic optimization. *ArXiv Preprint: 1802.03653*, 2018.

[Bihari(1956)] I. Bihari. A generalization of a lemma of Bellman and its application to uniqueness problems of differential equations. *Acta Mathematica Hungarica*, 7(1):81–94, 1956.

[Birgin et al.(2016)] E. G. Birgin, J. L. Gardenghi, J. M. Martinez, S. A. Santos, and P. L. Toint. Evaluation complexity for nonlinear constrained optimization using unscaled KKT conditions and high-order models. *SIAM Journal on Optimization*, 26(2):951–967, 2016.

[Birgin et al.(2017)] E. G. Birgin, J. L. Gardenghi, J. M. Martínez, S. A. Santos, and P. L. Toint. Worst-case evaluation complexity for unconstrained nonlinear optimization using high-order regularized models. *Mathematical Programming*, 163(1-2):359–368, 2017.

[Bolte et al.(2010)] J. Bolte, A. Daniilidis, O. Ley, and L. Mazet. Characterizations of łojasiewicz inequalities: subgradient flows, talweg, convexity. *Transactions of the American Mathematical Society*, 362 (6):3319–3363, 2010.

[Bot and Csetnek(2016)] R. I. Bot and E. R. Csetnek. Second order forward-backward dynamical systems for monotone inclusion problems. *SIAM Journal on Control and Optimization*, 54(3):1423–1443, 2016.

[Boţ et al.(2020)] R. I. Boţ, E. R. Csetnek, and S. C. László. Tikhonov regularization of a second order dynamical system with Hessian driven damping. *Mathematical Programming*, pages 1–36, 2020.

[Bubeck et al.(2019)] S. Bubeck, Q. Jiang, Y. T. Lee, Y. Li, and A. Sidford. Near-optimal method for highly smooth convex optimization. In *COLT*, pages 492–507. PMLR, 2019.

[Bullins(2020)] B. Bullins. Highly smooth minimization of nonsmooth problems. In *COLT*, pages 988–1030. PMLR, 2020.

[Bullins and Lai(2020)] B. Bullins and K. A. Lai. Higher-order methods for convex-concave min-max optimization and monotone variational inequalities. *ArXiv Preprint: 2007.04528*, 2020.

[Cartis et al.(2018)] C. Cartis, N. I. M. Gould, and P. L. Toint. Second-order optimality and beyond: Characterization and evaluation complexity in convexly constrained nonlinear optimization. *Foundations of Computational Mathematics*, 18(5):1073–1107, 2018.

[Cartis et al.(2019)] C. Cartis, N. I. Gould, and P. L. Toint. Universal regularization methods: Varying the power, the smoothness and the accuracy. *SIAM Journal on Optimization*, 29(1):595–615, 2019.

[Chergui(2008)] L. Chergui. Convergence of global and bounded solutions of a second order gradient like system with nonlinear dissipation and analytic nonlinearity. *Journal of Dynamics and Differential Equations*, 3(20):643–652, 2008.

[Chill and Fašangová(2010)] R. Chill and E. Fašangová. Gradient systems. In *Lecture Notes of the 13th International Internet Seminar, Matfyzpress, Prague*, 2010.

[Coddington and Levinson(1955)] E. A. Coddington and N. Levinson. *Theory of Ordinary Differential Equations*. Tata McGraw-Hill Education, 1955.

[Diakonikolas and Jordan(2020)] J. Diakonikolas and M. I. Jordan. Generalized momentum-based methods: A Hamiltonian perspective. *SIAM Journal on Optimization*, To appear, 2020.

[Diakonikolas and Orecchia(2019)] J. Diakonikolas and L. Orecchia. The approximate duality gap technique: A unified theory of first-order methods. *SIAM Journal on Optimization*, 29(1):660–689, 2019.

[Doikov and Nesterov(2019)] N. Doikov and Y. Nesterov. Local convergence of tensor methods. *Mathematical Programming*, pages 1–22, 2019.

[Fazlyab et al.(2018)] 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.

[França et al.(2020)] G. França, J. Sulam, D. P. Robinson, and R. Vidal. Conformal symplectic and relativistic optimization. *Journal of Statistical Mechanics: Theory and Experiment*, 2020(12):124008, 2020.

[França et al.(2021)] G. França, M. I. Jordan, and R. Vidal. On dissipative symplectic integration with applications to gradient-based optimization. *Journal of Statistical Mechanics: Theory and Experiment*, To appear, 2021.

[Gasnikov et al.(2019)] A. Gasnikov, P. Dvurechensky, E. Gorbunov, E. Vorontsova, D. Selikhanovych, and C. A. Uribe. Optimal tensor methods in smooth convex and uniformly convex optimization. In *COLT*, pages 1374–1391. PMLR, 2019.

[Granas and Dugundji(2013)] A. Granas and J. Dugundji. *Fixed Point Theory*. Springer Science & Business Media, 2013.

[Grapiglia and Nesterov(2017)] G. N. Grapiglia and Y. Nesterov. Regularized Newton methods for minimizing functions with Hölder continuous Hessians. *SIAM Journal on Optimization*, 27(1):478–506, 2017.

[Grapiglia and Nesterov(2019)] G. N. Grapiglia and Y. Nesterov. Accelerated regularized Newton methods for minimizing composite convex functions. *SIAM Journal on Optimization*, 29(1):77–99, 2019.

[Grapiglia and Nesterov(2020a)] G. N. Grapiglia and Y. Nesterov. Tensor methods for finding approximate stationary points of convex functions. *Optimization Methods and Software*, pages 1–34, 2020a.

[Grapiglia and Nesterov(2020b)] G. N. Grapiglia and Y. Nesterov. Tensor methods for minimizing convex functions with Hölder continuous higher-order derivatives. *SIAM Journal on Optimization*, 30(4): 2750–2779, 2020b.

[Güler(1992)] O. Güler. New proximal point algorithms for convex minimization. *SIAM Journal on Optimization*, 2(4):649–664, 1992.

[Hu and Lessard(2017)] B. Hu and L. Lessard. Dissipativity theory for Nesterov's accelerated method. In *ICML*, pages 1549–1557. JMLR. org, 2017.

[Huang(2006)] S-Z. Huang. *Gradient Inequalities: with Applications to Asymptotic Behavior and Stability of Gradient-like Systems*, volume 126. American Mathematical Soc., 2006.

[Jiang et al.(2019)] B. Jiang, H. Wang, and S. Zhang. An optimal high-order tensor method for convex optimization. In *COLT*, pages 1799–1801. PMLR, 2019.

[Jiang et al.(2020)] B. Jiang, T. Lin, and S. Zhang. A unified adaptive tensor approximation scheme to accelerate composite convex optimization. *SIAM Journal on Optimization*, 30(4):2897–2926, 2020.

[Kamzolov(2020)] D. Kamzolov. Near-optimal hyperfast second-order method for convex optimization. In *International Conference on Mathematical Optimization Theory and Operations Research*, pages 167–178. Springer, 2020.

[Kamzolov and Gasnikov(2020)] D. Kamzolov and A. Gasnikov. Near-optimal hyperfast second-order method for convex optimization and its sliding. *ArXiv Preprint: 2002.09050*, 2020.

[Krichene et al.(2015)] W. Krichene, A. Bayen, and P. L. Bartlett. Accelerated mirror descent in continuous and discrete time. In *NeurIPS*, pages 2845–2853, 2015.

[Kurdyka(1998)] K. Kurdyka. On gradients of functions definable in o-minimal structures. In *Annales de l'institut Fourier*, volume 48, pages 769–783, 1998.

[LaSalle(1949)] J. LaSalle. Uniqueness theorems and successive approximations. *Annals of Mathematics*, pages 722–730, 1949.

[Lessard et al.(2016)] L. Lessard, B. Recht, and A. Packard. Analysis and design of optimization algorithms via integral quadratic constraints. *SIAM Journal on Optimization*, 26(1):57–95, 2016.

[Maddison et al.(2018)] C. J. Maddison, D. Paulin, Y. W. Teh, B. O'Donoghue, and A. Doucet. Hamiltonian descent methods. *ArXiv Preprint: 1809.05042*, 2018.

[Maingé(2013)] P-E. Maingé. First-order continuous Newton-like systems for monotone inclusions. *SIAM Journal on Control and Optimization*, 51(2):1615–1638, 2013.

[Martinet(1970)] B. Martinet. Régularisation d'inéquations variationnelles par approximations successives. rev. française informat. *Recherche Opérationnelle*, 4:154–158, 1970.

[Martinet(1972)] B. Martinet. Détermination approchée d'un point fixe d'une application pseudo-contractante. *CR Acad. Sci. Paris*, 274(2):163–165, 1972.

[Martínez(2017)] J. Martínez. On high-order model regularization for constrained optimization. *SIAM Journal on Optimization*, 27(4):2447–2458, 2017.

[May(2017)] R. May. Asymptotic for a second-order evolution equation with convex potential and vanishing damping term. *Turkish Journal of Mathematics*, 41(3):681–685, 2017.

[Monteiro and Svaiter(2010)] R. D. C. Monteiro and B. F. Svaiter. On the complexity of the hybrid proximal extragradient method for the iterates and the ergodic mean. *SIAM Journal on Optimization*, 20(6): 2755–2787, 2010.

[Monteiro and Svaiter(2013)] R. D. C. Monteiro and B. F. Svaiter. An accelerated hybrid proximal extragradient method for convex optimization and its implications to second-order methods. *SIAM Journal on Optimization*, 23(2):1092–1125, 2013.

[Muehlebach and Jordan(2019)] M. Muehlebach and M. I. Jordan. A dynamical systems perspective on Nesterov acceleration. In *ICML*, pages 4656–4662, 2019.

[Muehlebach and Jordan(2021)] M. Muehlebach and M. I. Jordan. Optimization with momentum: Dynamical, control-theoretic, and symplectic perspectives. *The Journal of Machine Learning Research*, To appear, 2021.

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

[Nesterov(2012)] Y. Nesterov. How to make the gradients small. *Optima*, 88:10–11, 2012.

[Nesterov(2013)] Y. Nesterov. Gradient methods for minimizing composite functions. *Mathematical Programming*, 140(1):125–161, 2013.

[Nesterov(2018)] Y. Nesterov. *Lectures on Convex Optimization*, volume 137. Springer, 2018.

[Nesterov(2019)] Y. Nesterov. Implementable tensor methods in unconstrained convex optimization. *Mathematical Programming*, pages 1–27, 2019.

[Nesterov(2020a)] Y. Nesterov. Inexact accelerated high-order proximal-point methods. Technical report, Université catholique de Louvain, Center for Operations Research and Econometrics (CORE), 2020a.

[Nesterov(2020b)] Y. Nesterov. Superfast second-order methods for unconstrained convex optimization. Technical report, Université catholique de Louvain, Center for Operations Research and Econometrics (CORE), 2020b.

[Nesterov(1983)] Y. E. Nesterov. A method for solving the convex programming problem with convergence rate o (1/kˆ 2). In *Dokl. akad. nauk Sssr*, volume 269, pages 543–547, 1983.

[O'Donoghue and Maddison(2019)] B. O'Donoghue and C. J. Maddison. Hamiltonian descent for composite objectives. In *NeurIPS*, pages 14470–14480, 2019.

[Ostroukhov et al.(2020)] P. Ostroukhov, R. Kamalov, P. Dvurechensky, and A. Gasnikov. Tensor methods for strongly convex strongly concave saddle point problems and strongly monotone variational inequalities. *ArXiv Preprint: 2012.15595*, 2020.

[Polyak(1987)] B. T. Polyak. *Introduction to Optimization*. Optimization Software, Inc, New York, 1987.

[Rockafellar(1976)] R. T. Rockafellar. Monotone operators and the proximal point algorithm. *SIAM Journal on Control and Optimization*, 14(5):877–898, 1976.

[Scieur et al.(2017)] D. Scieur, V. Roulet, F. Bach, and A. d'Aspremont. Integration methods and optimization algorithms. In *NeurIPS*, pages 1109–1118, 2017.

[Sebbouh et al.(2020)] O. Sebbouh, C. Dossal, and A. Rondepierre. Convergence rates of damped inertial dynamics under geometric conditions and perturbations. *SIAM Journal on Optimization*, 30(3):1850–1877, 2020.

[Shi et al.(2018)] B. Shi, S. S. Du, M. I. Jordan, and W. J. Su. Understanding the acceleration phenomenon via high-resolution differential equations. *ArXiv Preprint: 1810.08907*, 2018.

[Shi et al.(2019)] B. Shi, S. S. Du, W. J. Su, and M. I. Jordan. Acceleration via symplectic discretization of high-resolution differential equations. In *NeurIPS*, pages 5744–5752, 2019.

[Solodov and Svaiter(1999)] M. V. Solodov and B. F. Svaiter. A hybrid approximate extragradient-proximal point algorithm using the enlargement of a maximal monotone operator. *Set-Valued Analysis*, 7(4): 323–345, 1999.

[Song et al.(2021)] C. Song, Y. Jiang, and Y. Ma. Unified acceleration of high-order algorithms under Hölder continuity and uniform convexity. *SIAM Journal on Optimization*, To appear, 2021.

[Su et al.(2016)] W. Su, S. Boyd, and E. J. Candès. A differential equation for modeling Nesterov's accelerated gradient method: theory and insights. *The Journal of Machine Learning Research*, 17(1): 5312–5354, 2016.

[Sutherland(2009)] W. A. Sutherland. *Introduction to Metric and Topological Spaces*. Oxford University Press, 2009.

[Tseng(2010)] P. Tseng. Approximation accuracy, gradient methods, and error bound for structured convex optimization. *Mathematical Programming*, 125(2):263–295, 2010.

[Vassilis et al.(2018)] A. Vassilis, A. Jean-François, and D. Charles. The differential inclusion modeling FISTA algorithm and optimality of convergence rate in the case b $\leq$ 3. *SIAM Journal on Optimization*, 28(1):551–574, 2018.

[Wibisono et al.(2016)] 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.

[Wilson et al.(2019)] A. C. Wilson, L. Mackey, and A. Wibisono. Accelerating rescaled gradient descent: Fast optimization of smooth functions. In *NeurIPS*, pages 13555–13565, 2019.

[Wilson et al.(2021)] A. C. Wilson, B. Recht, and M. I. Jordan. A Lyapunov analysis of momentum methods in optimization. *The Journal of Machine Learning Research*, To appear, 2021.

[Zhang et al.(2018)] J. Zhang, A. Mokhtari, S. Sra, and A. Jadbabaie. Direct Runge-Kutta discretization achieves acceleration. In *NeurIPS*, pages 3900–3909, 2018.

A Implicit Time Discretization and Optimal Acceleration

In this section, we propose two conceptual algorithmic frameworks that arise via implicit time discretization of the closed-loop system in Eq. (7) and Eq. (8). Our approach demonstrates the importance of the large-step condition [Monteiro and Svaiter(2013)] for optimal acceleration, interpreting it as the discretization of the algebraic equation. This allows us to further clarify why this condition is unnecessary for first-order optimization algorithms in the case of $p = 1$ (the algebraic equation disappears). With an approximate tensor subroutine [Nesterov(2019)], we derive two class of p-th order tensor algorithms, one of which recovers existing optimal p-th order tensor algorithms [Gasnikov et al.(2019), Jiang et al.(2019), Bubeck et al.(2019)] and the other of which leads to a new optimal p-th order tensor algorithm.

A.1 Conceptual algorithmic frameworks

We study two conceptual algorithmic frameworks which are derived by implicit time discretization of Eq. (7) with $c = 0$ and Eq. (8) with $c = 2$.

First algorithmic framework. By the definition of $a(t)$, we have $(\dot{a}(t))^2 = \lambda(t)a(t)$ and $a(0) = 0$. This implies an equivalent formulation of the first-order system in Eq. (7) with $c = 0$ as follows,

$$\left\{\begin{array}{l} \dot{v}(t) + \dot{a}(t)\nabla\Phi(x(t)) = 0 \\ \dot{x}(t) + \frac{\dot{a}(t)}{a(t)}(x(t) - v(t)) + \frac{(\dot{a}(t))^2}{a(t)}\nabla\Phi(x(t)) = 0 \\ a(t) = \frac{1}{4}(\int_0^t \sqrt{\lambda(s)}ds)^2 \\ (\lambda(t))^p \|\nabla\Phi(x(t))\|^{p-1} = \theta \\ (x(0), v(0)) = (x_0, v_0) \end{array}\right.$$

$$\Longleftrightarrow \left\{\begin{array}{l} \dot{v}(t) + \dot{a}(t)\nabla\Phi(x(t)) = 0 \\ a(t)\dot{x}(t) + \dot{a}(t)(x(t) - v(t)) + \lambda(t)a(t)\nabla\Phi(x(t)) = 0 \\ (\dot{a}(t))^2 = \lambda(t)a(t) \\ (\lambda(t))^p \|\nabla\Phi(x(t))\|^{p-1} = \theta \\ (x(0), v(0), a(0)) = (x_0, v_0, 0) \end{array}\right.$$

We define discrete-time sequences, $\{(x_k, v_k, \lambda_k, a_k, A_k)\}_{k\geq 0}$, that correspondx to the continuous-time sequences $\{(x(t), v(t), \lambda(t), \dot{a}(t), a(t))\}_{t\geq 0}$. By implicit time discretization, we have

$$\left\{\begin{array}{l} v_{k+1} - v_k + a_{k+1}\nabla\Phi(x_{k+1}) = 0 \\ A_{k+1}(x_{k+1} - x_k) + a_{k+1}(x_k - v_k) + \lambda_{k+1}A_{k+1}\nabla\Phi(x_{k+1}) = 0 \\ (a_{k+1})^2 = \lambda_{k+1}(A_k + a_{k+1}),\ a_{k+1} = A_{k+1} - A_k,\ a_0 = 0 \\ (\lambda_{k+1})^p \|\nabla\Phi(x_{k+1})\|^{p-1} = \theta \end{array}\right. \tag{14}$$

By introducing a new variable $\tilde{v}_k = \frac{A_k}{A_k + a_{k+1}}x_k + \frac{a_{k+1}}{A_k + a_{k+1}}v_k$, the second and fourth lines of Eq. (14) can be equivalently reformulated as follows:

$$\lambda_{k+1}\nabla\Phi(x_{k+1}) + x_{k+1} - \tilde{v}_k = 0, \qquad \lambda_{k+1}\|x_{k+1} - \tilde{v}_k\|^{p-1} = \theta$$

We propose to solve these two equations inexactly and replace $\nabla\Phi(x_{k+1})$ by a sufficiently accurate approximation in the first line of Eq. (14). In particular, the first equation can be equivalently written in the form of $\lambda_{k+1}w_{k+1} + x_{k+1} - \tilde{v}_k = 0$, where $w_{k+1} \in \{\nabla\Phi(x_{k+1})\}$. This motivates us to introduce a relative error tolerance [Solodov and Svaiter(1999), Monteiro and Svaiter(2010)]. In particular, we define the ε-subdifferential of a function f by

$$\partial_\epsilon f(x) := \{w \in \mathbb{R}^d \mid f(y) \geq f(x) + \langle y - x, w\rangle - \epsilon,\ \forall y \in \mathbb{R}^d\} \tag{15}$$

and find $\lambda_{k+1} > 0$ and a triple $(x_{k+1}, w_{k+1}, \varepsilon_{k+1})$ such that $\|\lambda_{k+1}w_{k+1} + x_{k+1} - \tilde{v}_k\|^2 + 2\lambda_{k+1}\epsilon_{k+1} \leq \sigma^2\|x_{k+1} - \tilde{v}_k\|^2$, where $w_{k+1} \in \partial_{\epsilon_{k+1}}\Phi(x_{k+1})$. To this end, w_{k+1} is a sufficiently accurate approximation of $\nabla\Phi(x_{k+1})$. Moreover, the second equation can be relaxed to $\lambda_{k+1}\|x_{k+1} - \tilde{v}_k\|^{p-1} \geq \theta$.

Remark. We present our first conceptual algorithmic framework formally in Algorithm 1. This scheme includes the large-step A-HPE framework [Monteiro and Svaiter(2013)] as a special instance. Indeed, it reduces to the large-step A-HPE framework if we set $y = \tilde{y}$ and $p = 2$ and change the notation of $(x, v, \tilde{v}, w)$ to $(y, x, \tilde{x}, v)$ in [Monteiro and Svaiter(2013)].

Algorithm 1 Conceptual Algorithmic Framework I

STEP 0: Let $x_0, v_0 \in \mathbb{R}^d$, $\sigma \in (0,1)$ and $\theta > 0$ be given, and set $A_0 = 0$ and $k = 0$
STEP 1: If $0 = \nabla\Phi(x_k)$, then **stop**
STEP 2: Otherwise, compute $\lambda_{k+1} > 0$ and a triple $(x_{k+1}, w_{k+1}, \epsilon_{k+1}) \in \mathbb{R}^d \times \mathbb{R}^d \times (0, +\infty)$ such that

$$\begin{aligned} &w_{k+1} \in \partial_{\epsilon_{k+1}}\Phi(x_{k+1}) \\ &\|\lambda_{k+1}w_{k+1} + x_{k+1} - \tilde{v}_k\|^2 + 2\lambda_{k+1}\epsilon_{k+1} \le \sigma^2\|x_{k+1} - \tilde{v}_k\|^2 \\ &\lambda_{k+1}\|x_{k+1} - \tilde{v}_k\|^{p-1} \ge \theta \end{aligned}$$

where $\tilde{v}_k = \frac{A_k}{A_k + a_{k+1}}x_k + \frac{a_{k+1}}{A_k + a_{k+1}}v_k$ and $a_{k+1}^2 = \lambda_{k+1}(A_k + a_{k+1})$
STEP 3: Compute $A_{k+1} = A_k + a_{k+1}$ and $v_{k+1} = v_k - a_{k+1}w_{k+1}$
STEP 4: Set $k \leftarrow k + 1$, and go to **STEP 1**

Second algorithmic framework. By the definition of $\gamma(t)$, we have $(\frac{\dot{\gamma}(t)}{\gamma(t)})^2 = \lambda(t)\gamma(t)$ and $\gamma(0) = 1$. This implies an equivalent formulation of the first-order system in Eq. (8) with $c = 2$:

$$\begin{cases} \dot{v}(t) - \frac{\dot{\gamma}(t)}{\gamma^2(t)}\nabla\Phi(x(t)) = 0 \\ \dot{x}(t) - \frac{\dot{\gamma}(t)}{\gamma(t)}(x(t) - v(t)) + \frac{(\dot{\gamma}(t))^2}{(\gamma(t))^3}\nabla\Phi(x(t)) = 0 \\ \gamma(t) = 4(\int_0^t \sqrt{\lambda(s)}ds + c)^{-2} \\ (\lambda(t))^p\|\nabla\Phi(x(t))\|^{p-1} = \theta \\ (x(0), v(0)) = (x_0, v_0) \end{cases}$$

$$\iff \begin{cases} \dot{v}(t) + \frac{\alpha(t)}{\gamma(t)}\nabla\Phi(x(t)) = 0 \\ \dot{x}(t) + \alpha(t)(x(t) - v(t)) + \lambda(t)\nabla\Phi(x(t)) = 0 \\ (\alpha(t))^2 = \lambda(t)\gamma(t),\ \dot{\gamma}(t) + \alpha(t)\gamma(t) = 0 \\ (\lambda(t))^p\|\nabla\Phi(x(t))\|^{p-1} = \theta \\ (x(0), v(0), \gamma(0)) = (x_0, v_0, 1) \end{cases}$$

We define discrete-time sequences, $\{(x_k, v_k, \lambda_k, \alpha_k, \gamma_k)\}_{k\ge 0}$, that correspondx to the continuous-time sequences $\{(x(t), v(t), \lambda(t), \alpha(t), \gamma(t))\}_{t\ge 0}$. From implicit time discretization, we have

$$\begin{cases} v_{k+1} - v_k + \frac{\alpha_{k+1}}{\gamma_{k+1}}\nabla\Phi(x_{k+1}) = 0 \\ x_{k+1} - x_k + \alpha_{k+1}(x_k - v_k) + \lambda_{k+1}\nabla\Phi(x_{k+1}) = 0 \\ (\alpha_{k+1})^2 = \lambda_{k+1}\gamma_{k+1},\ \gamma_{k+1} = (1 - \alpha_{k+1})\gamma_k,\ \gamma_0 = 1 \\ (\lambda_{k+1})^p\|\nabla\Phi(x_{k+1})\|^{p-1} = \theta \end{cases} \tag{16}$$

By introducing a new variable $\tilde{v}_k = (1 - \alpha_{k+1})x_k + \alpha_{k+1}v_k$, the second and fourth lines of Eq. (14) can be equivalently reformulated as

$$\lambda_{k+1}\nabla\Phi(x_{k+1}) + x_{k+1} - \tilde{v}_k = 0, \qquad \lambda_{k+1}\|x_{k+1} - \tilde{v}_k\|^{p-1} = \theta$$

By the same approximation strategy as before, we solve these two equations inexactly and replace $\nabla\Phi(x_{k+1})$ by a sufficiently accurate approximation in the first line of Eq. (16).

Remark. We present our second conceptual algorithmic framework formally in Algorithm 2. To the best of our knowledge, this scheme does not appear in the literature and is based on an

Algorithm 2 Conceptual Algorithmic Framework II

STEP 0: Let $x_0, v_0 \in \mathbb{R}^d$, $\sigma \in (0,1)$ and $\theta > 0$ be given, and set $\gamma_0 = 1$ and $k = 0$
STEP 1: If $0 = \nabla\Phi(x_k)$, then **stop**
STEP 2: Otherwise, compute $\lambda_{k+1} > 0$ and a triple $(x_{k+1}, w_{k+1}, \epsilon_{k+1}) \in \mathbb{R}^d \times \mathbb{R}^d \times (0, +\infty)$ such that

$$\begin{aligned} &w_{k+1} \in \partial_{\epsilon_{k+1}}\Phi(x_{k+1}) \\ &\|\lambda_{k+1}w_{k+1} + x_{k+1} - \tilde{v}_k\|^2 + 2\lambda_{k+1}\epsilon_{k+1} \leq \sigma^2\|x_{k+1} - \tilde{v}_k\|^2 \\ &\lambda_{k+1}\|x_{k+1} - \tilde{v}_k\|^{p-1} \geq \theta \end{aligned}$$

where $\tilde{v}_k = (1-\alpha_{k+1})x_k + \alpha_{k+1}v_k$ and $(\alpha_{k+1})^2 = \lambda_{k+1}(1-\alpha_{k+1})\gamma_k$
STEP 3: Compute $\gamma_{k+1} = (1-\alpha_{k+1})\gamma_k$ and $v_{k+1} = v_k - \frac{\alpha_{k+1}}{\gamma_{k+1}}w_{k+1}$
STEP 4: Set $k \leftarrow k+1$, and go to **STEP 1**

estimate sequence which differs from the one used in Algorithm 1. However, from a continuous-time perspective, these two algorithms are equivalent up to a constant $c > 0$, demonstrating that they achieve the same convergence rate in terms of both objective function gap and squared gradient norm.

Comparison with Güler's accelerated proximal point algorithm. Algorithm 2 is related to Güler's accelerated proximal point algorithm (APPA) [Güler(1992)], which combines Nesterov acceleration [Nesterov(1983)] and Martinet's PPA [Martinet(1970), Martinet(1972)]. Indeed, the analogs of update formulas $\tilde{v}_k = (1-\alpha_{k+1})x_k + \alpha_{k+1}v_k$ and $(\alpha_{k+1})^2 = \lambda_{k+1}(1-\alpha_{k+1})\gamma_k$ appear in Güler's algorithm, suggesting similar evolution dynamics. However, Güler's APPA does not specify how to choose $\{\lambda_k\}_{k\geq 0}$ but regard them as the parameters, while our algorithm links its choice with the gradient norm of Φ via the large-step condition.

Such difference is emphasized by recent studies on the continuous-time perspective of Güler's APPA [Attouch et al.(2019c), Attouch et al.(2019a)]. More specifically, [Attouch et al.(2019a)] proved that Güler's APPA can be interpreted as the implicit time discretization of an open-loop inertial gradient system (see [Attouch et al.(2019a), Eq. (53)]):

$$\ddot{x}(t) + \left(g(t) - \frac{\dot{g}(t)}{g(t)}\right)\dot{x}(t) + \beta(t)\nabla\Phi(x(t)) = 0$$

where g_k and β_k in their notation correspond to α_k and λ_k in Algorithm 2. By using $\gamma_{k+1} - \gamma_k = -\alpha_{k+1}\gamma_k$ and standard continuous-time arguments, we have $g(t) = -\frac{\dot{\gamma}(t)}{\gamma(t)}$ and $\beta(t) = \lambda(t) = \frac{(\dot{\gamma}(t))^2}{(\gamma(t))^3}$. By further defining $a(t) = \frac{1}{\gamma(t)}$, the above system is in the form of

$$\ddot{x}(t) + \left(\frac{2\dot{a}(t)}{a(t)} - \frac{\ddot{a}(t)}{\dot{a}(t)}\right)\dot{x}(t) + \left(\frac{(\dot{a}(t))^2}{a(t)}\right)\nabla\Phi(x(t)) = 0 \tag{17}$$

where a explicitly depends on the variable λ as follows,

$$a(t) = \tfrac{1}{4}(\textstyle\int_0^t \sqrt{\lambda(s)}ds + 2)^2$$

Compared to our closed-loop control system, the one in Eq. (17) is open-loop without the algebra equation and does not contain Hessian-driven damping. The coefficient for the gradient term is also different, standing for different time rescaling in the evolution dynamics [Attouch et al.(2021a)].

A.2 Complexity analysis

We study the iteration complexity of Algorithm 1 and 2. Our analysis is largely motivated by the aforementioned continuous-time analysis, simplifying the analysis in [Monteiro and Svaiter(2013)] for the case of $p = 2$ and generalizing it to the case of $p > 2$ in a systematic manner (see Theorem 3 and Theorem 4). Throughout this subsection, $x^\star$ denotes the projection of v_0 onto the solution set of Φ.

Algorithm 1. We start with the presentation of our main results for Algorithm 1, which in fact generalizes [Monteiro and Svaiter(2013), Theorem 4.1] to the case of $p > 2$.

Theorem 3. *For every integer $k \geq 1$, the objective function gap satisfies*

$$\Phi(x_k) - \Phi(x^\star) = O(k^{-\frac{3p+1}{2}})$$

and

$$\inf_{1\leq i\leq k} \|w_i\|^2 = O(k^{-3p}), \quad \inf_{1\leq i\leq k} \epsilon_i = O(k^{-\frac{3p+3}{2}})$$

Note that the only difference between Algorithm 1 and large-step A-HPE framework in [Monteiro and Svaiter(201 is the order in the algebraic equation. As such, many of the technical results derived in [Monteiro and Svaiter(2013)] also hold for Algorithm 1; more specifically, [Monteiro and Svaiter(2013), Theorem 3.6, Lemma 3.7 and Proposition 3.9].

Algorithm 2. We now present our main results for Algorithm 2. The proof is analogous to that of Theorem 3 and based on another estimate sequence.

Theorem 4. *For every integer $k \geq 1$, the objective function gap satisfies*

$$\Phi(x_k) - \Phi(x^\star) = O(k^{-\frac{3p+1}{2}})$$

and

$$\inf_{1\leq i\leq k} \|w_i\|^2 = O(k^{-3p}), \quad \inf_{1\leq i\leq k} \epsilon_i = O(k^{-\frac{3p+3}{2}})$$

A.3 Optimal tensor algorithms and gradient norm minimization

By instantiating Algorithm 1 and 2 with approximate tensor subroutines, we develop two families of optimal p-th order tensor algorithms for minimizing the function $\Phi \in \mathcal{F}_\ell^p(\mathbb{R}^d)$. The former one include all of existing optimal p-th order tensor algorithms [Gasnikov et al.(2019), Jiang et al.(2019), Bubeck et al.(2019)] while the latter one is new to our knowledge. We also provide one hitherto unknown result that the optimal p-th order tensor algorithms in this section minimize the squared gradient norm at a rate of $O(k^{-3p})$. The results extend those for the optimal first-order and second-order algorithms that have been obtained in [Shi et al.(2018)] and [Monteiro and Svaiter(2013)].

Approximate tensor subroutine. The celebrated proximal point algorithms [Rockafellar(1976), Güler(1992)] (corresponding to implicit time discretization of certain systems) require solving an exact proximal iteration with proximal coefficient $\lambda > 0$ at each iteration:

$$x = \operatorname*{argmin}_{u\in\mathbb{R}^d} \left\{ \Phi(u) + \frac{1}{2\lambda} \|u - v\|^2 \right\} \tag{18}$$

Algorithm 3 Optimal p-th order Tensor Algorithm I [Gasnikov et al.(2019), Jiang et al.(2019), Bubeck et al.(2019)]

STEP 0: Let $x_0, v_0 \in \mathbb{R}^d$, $\hat{\sigma} \in (0,1)$ and $0 < \sigma_l < \sigma_u < 1$ such that $\sigma_l(1+\hat{\sigma})^{p-1} < \sigma_u(1-\hat{\sigma})^{p-1}$ and $\sigma = \hat{\sigma} + \sigma_u < 1$ be given, and set $A_0 = 0$ and $k = 0$

STEP 1: If $0 = \nabla\Phi(x_k)$, then **stop**

STEP 2: Otherwise, compute a positive scalar λ_{k+1} with a $\hat{\sigma}$-inexact solution $x_{k+1} \in \mathbb{R}^d$ of Eq. (19a) satisfying that

$$\frac{\sigma_l p!}{2\ell} \leq \lambda_{k+1} \|x_{k+1} - \tilde{v}_k\|^{p-1} \leq \frac{\sigma_u p!}{2\ell}$$

or an exact solution $x_{k+1} \in \mathbb{R}^d$ of Eq. (19b) satisfying that

$$\frac{(p-1)!}{2\ell} \leq \lambda_{k+1} \|x_{k+1} - \tilde{v}_k\|^{p-1} \leq \frac{p!}{\ell(p+1)}$$

where $\tilde{v}_k = \frac{A_k}{A_k + a_{k+1}} x_k + \frac{a_{k+1}}{A_k + a_{k+1}} v_k$ and $a_{k+1}^2 = \lambda_{k+1}(A_k + a_{k+1})$

STEP 3: Compute $A_{k+1} = A_k + a_{k+1}$ and $v_{k+1} = v_k - a_{k+1}\nabla\Phi(x_{k+1})$

STEP 4: Set $k \leftarrow k+1$, and go to **STEP 1**

In general, Eq. (18) can be as hard as minimizing the function Φ when the proximal coefficient $\lambda \to +\infty$. Fortunately, when $\Phi \in \mathcal{F}_\ell^p(\mathbb{R}^d)$, it suffices to solve the subproblem that minimizes the sum of the p-th order Taylor approximation of Φ and a regularization term, motivating a line of p-th order tensor algorithms [Baes(2009), Birgin et al.(2016), Birgin et al.(2017), Martínez(2017), Nesterov(2019), Jiang et al.(2020), Gasnikov et al.(2019), Jiang et al.(2019), Bubeck et al.(2019)]. More specifically, we define

$$\Phi_v(u) = \Phi(v) + \langle \nabla\Phi(v), u - v\rangle + \sum_{j=2}^{p} \frac{1}{j!} \nabla^{(j)}\Phi(v)[u-v]^j + \frac{\ell\|u-v\|^{p+1}}{(p+1)!}$$

The algorithms of this subsection are based on either an inexact solution of Eq. (19a), used in [Jiang et al.(2019)], or an exact solution of Eq. (19b), used in [Gasnikov et al.(2019)] and [Bubeck et al.(2019)]:

$$\min_{u \in \mathbb{R}^d} \ \Phi_v(u) + \frac{1}{2\lambda}\|u - v\|^2 \tag{19a}$$

$$\min_{u \in \mathbb{R}^d} \ \Phi_v(u) \tag{19b}$$

In particular, the solution x_v of Eq. (19a) is unique and satisfies $\lambda\nabla\Phi_v(x_v) + x_v - v = 0$. Thus, we denote a *$\hat{\sigma}$-inexact solution* of Eq. (19a) by a vector $x \in \mathbb{R}^d$ satisfying that $\|\lambda\nabla\Phi_v(x) + x - v\| \leq \hat{\sigma}\|x - v\|$ use either it or an exact solution of Eq. (19b) in our tensor algorithms.

First algorithm. We present the first optimal p-th order tensor algorithm in Algorithm 3 and prove that it is Algorithm 1 with specific choice of θ.

Theorem 5. *Algorithm 3 is Algorithm 1 with $\theta = \frac{\sigma_l p!}{2\ell}$ or $\theta = \frac{(p-1)!}{2\ell}$.*

In view of Theorem 5, the iteration complexity derived for Algorithm 1 hold for Algorithm 3. We summarize the results in the following theorem.

Algorithm 4 Optimal p-th order Tensor Algorithm II

STEP 0: Let $x_0, v_0 \in \mathbb{R}^d$, $\hat{\sigma} \in (0,1)$ and $0 < \sigma_l < \sigma_u < 1$ such that $\sigma_l(1+\hat{\sigma})^{p-1} < \sigma_u(1-\hat{\sigma})^{p-1}$ and $\sigma = \hat{\sigma} + \sigma_u < 1$ be given, and set $\gamma_0 = 1$ and $k = 0$

STEP 1: If $0 = \nabla\Phi(x_k)$, then **stop**

STEP 2: Otherwise, compute a positive scalar λ_{k+1} with a $\hat{\sigma}$-inexact solution $x_{k+1} \in \mathbb{R}^d$ of Eq. (19a) satisfying that

$$\frac{\sigma_l p!}{2\ell} \leq \lambda_{k+1}\|x_{k+1} - \tilde{v}_k\|^{p-1} \leq \frac{\sigma_u p!}{2\ell}$$

or an exact solution $x_{k+1} \in \mathbb{R}^d$ of Eq. (19b) satisfying that

$$\frac{(p-1)!}{2\ell} \leq \lambda_{k+1}\|x_{k+1} - \tilde{v}_k\|^{p-1} \leq \frac{p!}{\ell(p+1)}$$

where $\tilde{v}_k = (1-\alpha_{k+1})x_k + \alpha_{k+1}v_k$ and $(\alpha_{k+1})^2 = \lambda_{k+1}(1-\alpha_{k+1})\gamma_k$

STEP 3: Compute $\gamma_{k+1} = (1-\alpha_{k+1})\gamma_k$ and $v_{k+1} = v_k - \frac{\alpha_{k+1}\nabla\Phi(x_{k+1})}{\gamma_{k+1}}$

STEP 4: Set $k \leftarrow k+1$, and go to **STEP 1**

Theorem 6 (Proof omitted)**.** *For every integer $k \geq 1$, the objective function gap satisfies*

$$\Phi(x_k) - \Phi(x^\star) = O(k^{-\frac{3p+1}{2}})$$

and the squared gradient norm satisfies

$$\inf_{1\leq i\leq k} \|\nabla\Phi(x_i)\|^2 = O(k^{-3p})$$

Remark. Theorem 6 has been derived in [Monteiro and Svaiter(2013), Theorem 6.4] for the special case of $p = 2$, and a similar result for Nesterov's accelerated gradient descent (the special case of $p = 1$) has also been derived in [Shi et al.(2018)]. For $p \geq 3$ in general, the first inequality on the objective function gap has been derived independently in [Gasnikov et al.(2019), Theorem 1], [Jiang et al.(2019), Theorem 3.5] and [Bubeck et al.(2019), Theorem 1.1], while the second inequality on the squared gradient norm is new to our knowledge.

Second algorithm. We present the second optimal p-th order tensor algorithm in Algorithm 4 which is Algorithm 2 with specific choice of θ. The proof is omitted since it is the same as the aforementioned analysis for Algorithm 3.

Theorem 7 (Proof omitted)**.** *Algorithm 4 is Algorithm 2 with $\theta = \frac{\sigma_l p!}{2\ell}$ or $\theta = \frac{(p-1)!}{2\ell}$.*

Theorem 8 (Proof omitted)**.** *For every integer $k \geq 1$, the objective gap satisfies*

$$\Phi(x_k) - \Phi(x^\star) = O(k^{-\frac{3p+1}{2}})$$

and the squared gradient norm satisfies

$$\inf_{1\leq i\leq k} \|\nabla\Phi(x_i)\|^2 = O(k^{-3p})$$

Remark. The approximate tensor subroutine in Algorithm 3 and 4 can be efficiently implemented usinga novel bisection search scheme. We refer the interested readers to [Jiang et al.(2019)] and [Bubeck et al.(2019)] for the details.

B Proofs for Section 3

B.1 Proof of Proposition 1

Proof of Proposition 1. By the definition of φ, the mapping $\varphi(\cdot, x)$ is linear and $\varphi(0, x) = 0$. Since $\nabla\Phi(x) \neq 0$, we have $\|\nabla\Phi(x)\| > 0$ and $\varphi(\cdot, x)$ is thus strictly increasing. Since $\varphi(\cdot, x)$ is linear and strictly increasing, $\varphi(\lambda, x) \to +\infty$ as $\lambda \to +\infty$. □

B.2 Proof of Proposition 2

Proof of Proposition 2. Given $x \in \Omega$, it suffices to show that $\mathbb{B}_\delta(x) \subseteq \Omega$ for some $\delta > 0$. Since Φ is twice continuously differentiable, $\nabla\Phi$ is locally Lipschitz; that is, there exists $\tilde{\delta} > 0$ and $L > 0$ such that

$$\|\nabla\Phi(z) - \nabla\Phi(x)\| \leq L\|z - x\|, \quad \forall z \in \mathbb{B}_{\delta_1}(x)$$

Combining this inequality with the triangle inequality, we have

$$\|\nabla\Phi(z)\| = \|\nabla\Phi(x)\| - \|\nabla\Phi(z) - \nabla\Phi(x)\| \geq \|\nabla\Phi(x)\| - L\|z - x\|$$

Let $\delta = \min\{\tilde{\delta}, \frac{\|\nabla\Phi(x)\|}{2L}\}$. Then, for any $z \in \mathbb{B}_\delta(x)$, we have

$$\|\nabla\Phi(z)\| \geq \frac{\|\nabla\Phi(x)\|}{2} > 0 \implies z \in \Omega$$

This completes the proof. □

B.3 Proof of Proposition 3

Proof of Proposition 3. By the definition of $\Lambda_\theta(\cdot)$, it suffices to show that $\Lambda_\theta(\cdot)$ is continuous and locally Lipschitz over Ω since the same argument works for $\sqrt{\Lambda_\theta(\cdot)}$.

First, we prove the continuity of $\Lambda_\theta(\cdot)$ over Ω. Since $\|\nabla\Phi(x)\| > 0$ for any $x \in \Omega$, the function $\|\nabla\Phi(\cdot)\|^{-\frac{p-1}{p}}$ is continuous over Ω. By the definition of $\Lambda_\theta(\cdot)$, we achieve the desired result.

Second, we prove that $\Lambda_\theta(\cdot)$ is locally Lipschitz over Ω. Since Φ is twice continuously differentiable, $\nabla\Phi$ is locally Lipschitz. For $p = 1$, $\Lambda_\theta(\cdot)$ is a constant everywhere and thus locally Lipschitz over Ω. For $p \geq 2$, the function $x^{-\frac{p-1}{p}}$ is locally Lipschitz at any point $x > 0$. Also, by Proposition 2, Ω is an open set. Putting these pieces together yields that $\|\nabla\Phi(\cdot)\|^{-\frac{p-1}{p}}$ is locally Lipschitz over Ω; that is, there exist $\delta > 0$ and $L > 0$ such that

$$|\|\nabla\Phi(x')\|^{-\frac{p-1}{p}} - \|\nabla\Phi(x'')\|^{-\frac{p-1}{p}}| \leq L\|x' - x''\|, \quad \forall x', x'' \in \mathbb{B}_\delta(x)$$

which implies that

$$|\Lambda_\theta(x') - \Lambda_\theta(x'')| \leq \theta^{\frac{1}{p}} L\|x' - x''\|, \quad \forall x', x'' \in \mathbb{B}_\delta(x)$$

This completes the proof. □

B.4 Proof of Proposition 4

Proof of Proposition 4. By Proposition 2 and the initial condition $x_0 \in \Omega$, there exists $\delta > 0$ such that $\mathbb{B}_\delta(x_0) \subseteq \Omega$. Note that Φ is twice continuously differentiable. By the definition of Λ_θ, we

obtain that $\Lambda_\theta(z)$ and $\nabla\Phi(z)$ are both bounded for any $z \in \mathbb{B}_\delta(x_0)$. Putting these pieces together shows that there exists $M > 0$ such that, for any continuous function $x : [0,1] \mapsto \mathbb{B}_\delta(x_0)$, we have

$$\|F(t, x(t))\| \leq M, \quad \forall t \in [0,1] \tag{20}$$

The set of such functions is not empty since a constant function $x = x_0$ is one element. Letting $t_1 = \min\{1, \frac{\delta}{M}\}$, we define $\mathcal{X}$ as the space of all continuous functions x on $[0, t_0]$ for some $t_0 < t_1$ whose graph is contained entirely inside the rectangle $[0, t_0] \times \mathbb{B}_\delta(x_0)$. For any $x \in \mathcal{X}$, we define

$$z(t) = Tx = x_0 + \int_0^t F(s, x(s))ds$$

Note that $z(\cdot)$ is well defined and continuous on $[0, t_0]$. Indeed, $x \in \mathcal{X}$ implies that $x(t) \in \mathbb{B}_\delta(x_0) \subseteq \Omega$ for $\forall t \in [0, t_0]$. Thus, the integral of $F(s, x(s))$ is well defined and continuous. Second, the graph of $z(t)$ lies entirely inside the rectangle $[0, t_0] \times \mathbb{B}_\delta(x_0)$. Indeed, since $t \leq t_0 < t_1 = \min\{1, \frac{\delta}{M}\}$, we have

$$\|z(t) - x_0\| = \left\| \int_0^t F(s, x(s))ds \right\| \overset{\text{Eq. (20)}}{\leq} Mt \leq Mt_0 \leq Mt_1 \leq \delta$$

Putting these pieces together yields that T maps $\mathcal{X}$ to itself. By the fundamental theorem of calculus, we have $\dot{z}(t) = F(t, x(t))$. By a standard argument from ordinary differential equation theory, $\dot{x}(t) = F(t, x(t))$ and $x(0) = x_0$ if and only if x is a fixed point of T. Thus, it suffices to show the existence and uniqueness of a fixed point of T.

We consider the Picard iterates $\{\psi_k\}_{k\geq 0}$ with $\psi_0(t) = x_0$ for $\forall t \in [0, t_0]$ and $\psi_{k+1} = T\psi_k$ for all $k \geq 0$. By the Banach fixed-point theorem [Granas and Dugundji(2013)], the Picard iterates converge to a unique fixed point of T if $\mathcal{X}$ is an nonempty and complete metric space and T is a contraction from $\mathcal{X}$ to $\mathcal{X}$.

First, we show that $\mathcal{X}$ is an nonempty and complete metric space. Indeed, we define $d(x, x') = \max_{t\in[0,t_0]} \|x(t) - x'(t)\|$. It is easy to verify that d is a metric and $(\mathcal{X}, d)$ is a complete metric space (see [Sutherland(2009)] for the details). In addition, $\mathcal{X}$ is nonempty since the constant function $x = x_0$ is one element.

It remains to prove that T is a contraction for some $t_0 < t_1$. Indeed, $\Lambda_\theta(z)$ and $\nabla\Phi(z)$ are bounded for $\forall z \in \mathbb{B}_\delta(x_0)$; that is, there exists $M_1 > 0$ such that $\max\{\Lambda_\theta(z), \|\nabla\Phi(z)\|\} \leq M_1$ for $\forall z \in \mathbb{B}_\delta(x_0)$. By Proposition 3, Λ_θ and $\sqrt{\Lambda_\theta}$ are continuous and locally Lipschitz over Ω. Since $\mathbb{B}_\delta(x_0) \subseteq \Omega$ is bounded, there exists $L_1 > 0$ such that, for any $x', x'' \in \mathbb{B}_\delta(x_0)$, we have

$$\max\{|\Lambda_\theta(x') - \Lambda_\theta(x'')|, |\sqrt{\Lambda_\theta}(x') - \sqrt{\Lambda_\theta}(x'')|\} \leq L_1\|x' - x''\| \tag{21}$$

Note that Φ is twice continuously differentiable. Thus, there exists $L_2 > 0$ such that $\|\nabla\Phi(x') - \nabla\Phi(x'')\| \leq L_2\|x' - x''\|$ for $\forall x', x'' \in \mathbb{B}_\delta(x_0)$. In addition, for any $t \in [0, t_0]$, we have $\|x(t)\| \leq \|x_0\| + \delta = M_2$.

We now proceed to the main proof. By the triangle inequality, we have

$$\|Tx'(t) - Tx''(t)\| \leq \underbrace{\int_0^t \|\Lambda_\theta(x'(s))\nabla\Phi(x'(s)) - \Lambda_\theta(x''(s))\nabla\Phi(x''(s))\|ds}_{\mathbf{I}}$$
$$+ \underbrace{\int_0^t \left\| \frac{\sqrt{\Lambda_\theta(x'(s))}}{\int_0^s \sqrt{\Lambda_\theta(x'(w))}dw + c} \left(\int_0^s \left(\sqrt{\Lambda_\theta(x'(w))} \left(\int_0^w \sqrt{\Lambda_\theta(x'(v))}\, dv + c \right) \right) \nabla\Phi(x'(w))dw \right) \right.}_{\mathbf{II}}$$
$$\underbrace{\left. - \frac{\sqrt{\Lambda_\theta(x''(s))}}{\int_0^s \sqrt{\Lambda_\theta(x''(w))}dw + c} \left(\int_0^s \left(\sqrt{\Lambda_\theta(x''(w))} \left(\int_0^w \sqrt{\Lambda_\theta(x''(v))}\, dv + c \right) \right) \nabla\Phi(x''(w))dw \right) \right\| ds}_{\mathbf{II}}$$
$$+ \underbrace{\int_0^t \left\| \frac{2\sqrt{\Lambda_\theta(x'(s))}}{\int_0^s \sqrt{\Lambda_\theta(x'(w))}dw + c}(x'(s) - v_0) - \frac{2\sqrt{\Lambda_\theta(x''(s))}}{\int_0^s \sqrt{\Lambda_\theta(x''(w))}dw + c}(x''(s) - v_0) \right\| ds}_{\mathbf{III}}$$

The key inequality for the subsequent analysis is as follows:

$$\|a_1 b_1 - a_2 b_2\| \leq \|a_1\|\|b_1 - b_2\| + \|b_2\|\|a_1 - a_2\| \tag{22}$$

First, by combining Eq. (22) with $\max\{\Lambda_\theta(x(t)), \|\nabla\Phi(x(t))\|\} \leq M_1$, $\|\nabla\Phi(x') - \nabla\Phi(x'')\| \leq L_2\|x' - x''\|$ and Eq. (21), we obtain:

$$\mathbf{I} \leq M_1(L_1 + L_2)t_0 d(x', x'')$$

Second, we combine Eq. (22) with $\sqrt{\Lambda_\theta(x(t))} \leq \sqrt{M_1}$, Eq. (21) and $0 < s \leq t_0 < t_1 < 1$ to obtain:

$$\left\| \frac{\sqrt{\Lambda_\theta(x'(s))}}{\int_0^s \sqrt{\Lambda_\theta(x'(w))}dw + c} - \frac{\sqrt{\Lambda_\theta(x''(s))}}{\int_0^s \sqrt{\Lambda_\theta(x''(w))}dw + c} \right\| \leq \left(\frac{1}{c} + \frac{2\sqrt{M_1}}{c^2} \right) L_1 d(x', x'')$$

We also obtain by combining Eq. (22) with $\max\{\Lambda_\theta(x(t)), \|\nabla\Phi(x(t))\|\} \leq M_1$, $\|\nabla\Phi(x') - \nabla\Phi(x'')\| \leq L_2\|x' - x''\|$, Eq. (21) and $0 < w \leq s \leq t_0 < t_1 < 1$ that

$$\begin{aligned}
&\left\| \int_0^s \left(\sqrt{\Lambda_\theta(x'(w))} \left(\int_0^w \sqrt{\Lambda_\theta(x'(v))}\, dv + c \right) \right) \nabla\Phi(x'(w))dw \right. \\
&\quad \left. - \int_0^s \left(\sqrt{\Lambda_\theta(x''(w))} \left(\int_0^w \sqrt{\Lambda_\theta(x''(v))}\, dv + c \right) \right) \nabla\Phi(x''(w))dw \right\| \\
&\leq (M_1 L_2 + c\sqrt{M_1}L_2 + 2(M_1)^{3/2}L_1 + cM_1 L_1) d(x', x'')
\end{aligned}$$

In addition, by using $\max\{\Lambda_\theta(x(t)), \|\nabla\Phi(x(t))\|\} \leq M_1$ and $0 < w \leq s \leq t_0 < t_1 < 1$, we have

$$\left\| \frac{\sqrt{\Lambda_\theta(x'(s))}}{\int_0^s \sqrt{\Lambda_\theta(x'(w))}\, dw + c} \right\| \leq \frac{\sqrt{M_1}}{c}$$
$$\left\| \int_0^s \left(\sqrt{\Lambda_\theta(x''(w))} \left(\int_0^w \sqrt{\Lambda_\theta(x''(v))}\, dv + c \right) \right) \nabla\Phi(x''(w))dw \right\| \leq (M_1)^2 + c(M_1)^{3/2}$$

Putting these pieces together yields that

$$\mathbf{II} \leq \left(\frac{2(M_1)^{5/2}L_1}{c^2} + \frac{(M_1)^{3/2}L_2 + 5(M_1)^2 L_1}{c} + M_1 L_2 + 2(M_1)^{3/2}L_1 \right) t_0 d(x', x'')$$

Finally, by a similar argument, we have

$$\mathbf{III} \leq \left(\frac{2\sqrt{M_1} + 2(M_2 + \|v_0\|)L_1}{c} + \frac{4\sqrt{M_1}(M_2 + \|v_0\|)L_1}{c^2} \right) t_0 d(x', x'')$$

Combining the upper bounds for **I**, **II** and **III**, we have

$$d(Tx', Tx'') = \max_{t \in [0,t_0]} \|Tx'(t) - Tx''(t)\| \leq \bar{M} t_0 d(x', x'')$$

where $\bar{M}$ is a constant that does not depend on t_0 (in fact it depends on c, x_0, δ, $\Phi(\cdot)$ and $\Lambda_\theta(\cdot)$) and is defined as follows:

$$\begin{aligned} \bar{M} \;=\; & \frac{2((M_1)^2 + 2M_2 + 2\|v_0\|)\sqrt{M_1}L_1}{c^2} + \frac{2\sqrt{M_1} + (2M_2 + 2\|v_0\| + 5(M_1)^2)L_1 + (M_1)^{3/2}L_2}{c} \\ & + 2M_1 L_2 + (M_1 + 2(M_1)^{3/2})L_1 \end{aligned}$$

Therefore, the mapping T is a contraction if $t_0 \in (0, t_1]$ satisfies $t_0 \leq \frac{1}{2\bar{M}}$. This completes the proof. □

C Discussion for Section 3

We compare the closed-loop control system in Eq. (3) and Eq. (4) with four main classes of systems in the literature.

Hessian-driven damping. The formal introduction of Hessian-driven damping in optimization dates to [Alvarez et al.(2002)], with many subsequent developments; see, e.g., [Attouch et al.(2016b)]. The system studied in this literature takes the following form:

$$\ddot{x}(t) + \frac{\alpha}{t}\dot{x}(t) + \beta \nabla^2 \Phi(x(t))\dot{x}(t) + \nabla \Phi(x(t)) = 0$$

In a Hilbert space setting and when $\alpha > 3$, the literature has established the weak convergence of any solution trajectory to a global minimizer of Φ and the convergence rate of $o(1/t^2)$ in terms of objective function gap.

Recall also that [Shi et al.(2018)] interpreted Nesterov acceleration as the discretization of a high-resolution differential equation:

$$\ddot{x}(t) + \frac{3}{t}\dot{x}(t) + \sqrt{s}\nabla^2 \Phi(x(t))\dot{x}(t) + \left(1 + \frac{3\sqrt{s}}{2t}\right) \nabla \Phi(x(t)) = 0$$

and showed that this equation distinguishes between Polyak's heavy-ball method and Nesterov's accelerated gradient method. In the special case in which $c = 0$ and $p = 1$, our system in Eq. (3) and Eq. (4) becomes

$$\ddot{x}(t) + \frac{3}{t}\dot{x}(t) + \theta \nabla^2 \Phi(x(t))\dot{x}(t) + \left(\theta + \frac{\theta}{t}\right) \nabla \Phi(x(t)) = 0 \tag{23}$$

which also belongs to the class of high-resolution differential equations. Moreover, for $c = 0$ and $p = 1$, our system can be studied within the recently-proposed framework of [Attouch et al.(2020a), Attouch et al.(2021a)]; indeed, in this case (α, β, b) in [Attouch et al.(2021a), Theorem 2.1] has an analytic form. However, the choice of (α, β, b) in our general setting in Eq. (4), for $p \geq 2$, does not have an analytic form and it is difficult to verify whether (α, β, b) in this case satisfies their condition.

Newton and Levenberg-Marquardt regularized systems. The precursor of this perspective was developed by [Alvarez and Pérez C(1998)] in a variational characterization of general regularization algorithms. By constructing the regularization of the potential function $\Phi(\cdot, \epsilon)$ satisfying $\Phi(\cdot, \epsilon) \to \Phi$ as $\epsilon \to 0$, they studied the following system:

$$\nabla^2 \Phi(x(t), \epsilon(t))\dot{x}(t) + \dot{\epsilon}(t)\frac{\partial^2 \Phi}{\partial \epsilon \partial x}(x(t), \epsilon(t)) + \nabla \Phi(x(t), \epsilon(t)) = 0$$

Subsequently, [Attouch and Redont(2001)] and [Attouch and Svaiter(2011)] studied Newton dissipative and Levenberg-Marquardt regularized systems:

$$\textbf{(Newton)} \qquad \ddot{x}(t) + \nabla^2 \Phi(x(t))\dot{x}(t) + \nabla \Phi(x(t)) = 0.$$
$$\textbf{(Levenberg-Marquardt)} \qquad \lambda(t)\dot{x}(t) + \nabla^2 \Phi(x(t))\dot{x}(t) + \nabla \Phi(x(t)) = 0$$

These systems have been shown to be well defined and stable with robust asymptotic behavior [Attouch and Svaiter(2011), Attouch et al.(2013), Abbas et al.(2014)], further motivating the study of the following inertial gradient system with constant damping and Hessian-driven damping [Alvarez et al.(2002)]:

$$\ddot{x}(t) + \alpha \dot{x}(t) + \beta \nabla^2 \Phi(x(t))\dot{x}(t) + \nabla \Phi(x(t)) = 0$$

This system attains strong asymptotic stabilization and fast convergence properties [Alvarez et al.(2002), Attouch et al.(2012)] and can be extended to solve the monotone inclusion problems with theoretical guarantee [Attouch and Svaiter(2011), Maingé(2013), Attouch et al.(2013), Abbas et al.(2014), Attouch et al.(2016a), Attouch and László(2020b), Attouch and László(2020a)]. However, all of these systems are aimed at interpreting standard and regularized Newton algorithms and fail to model optimal acceleration even for the second-order algorithms in [Monteiro and Svaiter(2013)].

Recently, [Attouch et al.(2016a)] proposed a proximal Newton algorithm for solving monotone inclusions, which is motivated by a closed-loop control system without inertia. This algorithm attains a suboptimal convergence rate of $O(t^{-2})$ in terms of objective function gap.

Closed-loop control systems. The closed-loop damping approach in [Attouch et al.(2013), Attouch et al.(2016a)] closely resembles ours. In particular, they interpret various Newton-type methods as the discretization of the closed-loop control system without inertia and prove the existence and uniqueness of a solution as well as the convergence rate of the solution trajectory. There are, however, some significant differences between our work and theirs. In particular, the appearance of inertia is well known to make analysis much more challenging. Standard existence and uniqueness proofs based on the Cauchy-Schwarz theorem suffice to analyze the system of [Attouch et al.(2013), Attouch et al.(2016a)] thanks to the lack of inertia, while Picard iterates and the Banach fixed-point theorem are necessary for our analysis. The construction of the Lyapunov function is also more difficult for the system with inertia.

This is an active research area and we refer the interested reader to a recent article of [Attouch et al.(2020a)] for a comprehensive treatment of this topic.

Continuous-time interpretation of high-order tensor algorithms. There is comparatively little work on continuous-time perspectives on high-order tensor algorithms; indeed, we are aware of only [Wibisono et al.(2016)] and [Song et al.(2021)].

By appealing to a variational formulation, [Wibisono et al.(2016)] derived the following inertial gradient system with asymptotic vanishing damping:

$$\ddot{x}(t) + \frac{p+2}{t}\dot{x}(t) + C(p+1)^2 t^{p-1}\nabla\Phi(x(t)) = 0 \tag{24}$$

Compared to our closed-loop control system, in Eq. (3) and Eq. (4), the system in Eq. (24) is an open-loop system without the algebra equation and does not contain Hessian-driven damping. These differences yield solution trajectories that only attain a suboptimal convergence rate of $O(t^{-(p+1)})$ in terms of objective function gap.

Very recently, [Song et al.(2021)] have proposed and analyzed the following dynamics (we consider the Euclidean setting for simplicity):

$$\begin{cases} a(t)\dot{x}(t) = \dot{a}(t)(z(t) - x(t)) \\ z(t) = \operatorname{argmin}_{x\in\mathbb{R}^d} \int_0^t \dot{a}(s)(\Phi(x(s)) + \langle\nabla\Phi(x(s)), x - x(s)\rangle)ds + \frac{1}{2}\|x - x_0\|^2 \end{cases}$$

Solving the minimization problem yields $z(t) = x_0 - \int_0^t \dot{a}(s)\nabla\Phi(x(s))ds$. Substituting and rearranging yields:

$$\ddot{x}(t) + \left(\frac{2\dot{a}(t)}{a(t)} - \frac{\ddot{a}(t)}{\dot{a}(t)}\right)\dot{x}(t) + \left(\frac{(\dot{a}(t))^2}{a(t)}\right)\nabla\Phi(x(t)) = 0 \tag{25}$$

Compared to our closed-loop control system, the system in (25) is open-loop and lacks Hessian-driven damping. Moreover, $a(t)$ needs to be determined by hand and [Song et al.(2021)] do not establish existence or uniqueness of solutions.

D Proofs for Section 4

D.1 Proof of Theorem 1

We provide two technical lemmas that characterize the descent property of $\mathcal{E}$ and the boundedness of the local solution $(x, v) : [0, t_0] \mapsto \mathbb{R}^d \times \mathbb{R}^d$.

Lemma 1. *Suppose that $(x, v, \lambda, a) : [0, t_0] \mapsto \mathbb{R}^d \times \mathbb{R}^d \times (0, +\infty) \times (0, +\infty)$ is a local solution of the first-order system in Eq. (7). Then, we have*

$$\frac{d\mathcal{E}(t)}{dt} \leq -a(t)\theta^{\frac{1}{p}}\|\nabla\Phi(x(t))\|^{\frac{p+1}{p}}, \quad \forall t \in [0, t_0]$$

Lemma 2. *Suppose that $(x, v, \lambda, a) : [0, t_0] \mapsto \mathbb{R}^d \times \mathbb{R}^d \times (0, +\infty) \times (0, +\infty)$ is a local solution of the first-order system in Eq. (7). Then, $(x(\cdot), v(\cdot))$ is bounded over the interval $[0, t_0]$ and the upper bound only depends on the initial condition.*

Proof of Theorem 1. We are ready to prove our main result on the existence and uniqueness of a global solution. In particular, let us consider a maximal solution of the closed-loop control system in Eq. (3) and Eq. (4):

$$(x, \lambda, a) : [0, T_{\max}) \mapsto \Omega \times (0, +\infty) \times (0, +\infty)$$

The existence of a maximal solution follows from a classical argument relying on the existence and uniqueness of a local solution (see Proposition 4).

It remains to show that the maximal solution is a global solution; that is, $T_{\max} = +\infty$, if λ is absolutely continuous on any finite bounded interval. Indeed, the property of λ guarantees that $\lambda(\cdot)$ is bounded on the interval $[0, T_{\max})$. By Lemma 2 and the equivalence between the closed-loop control system in Eq. (3) and Eq. (4) and the first-order system in Eq. (7), the solution trajectory $x(\cdot)$ is bounded on the interval $[0, T_{\max})$ and the upper bound only depends on the initial condition. This implies that $\dot{x}(\cdot)$ is also bounded on the interval $[0, T_{\max})$ by considering the system in the autonomous form of Eq. (11) and (12). Putting these pieces together yields that $x(\cdot)$ is Lipschitz continuous on $[0, T_{\max})$ and there exists $\bar{x} = \lim_{t \to T_{\max}} x(t)$.

If $T_{\max} < +\infty$, the absolute continuity of λ on any finite bounded interval implies that $\lambda(\cdot)$ is bounded on $[0, T_{\max}]$. This together with the algebraic equation implies that $\bar{x} \in \Omega$. However, by Proposition 4 with initial data $\bar{x}$, we can extend the solution to a strictly larger interval which contradicts the maximality of the aforementioned solution. This completes the proof. □

D.2 Proof of Theorem 2

The following lemma is a global version of Lemma 1 and the proof is exactly the same. Thus, we only state the result.

Lemma 3 (Proof omitted)**.** *Suppose that $(x, v, \lambda, a) : [0, +\infty) \mapsto \mathbb{R}^d \times \mathbb{R}^d \times (0, +\infty) \times (0, +\infty)$ is a global solution of the first-order system in Eq. (7). Then, we have*

$$\frac{d\mathcal{E}(t)}{dt} \leq -a(t)\theta^{\frac{1}{p}} \|\nabla \Phi(x(t))\|^{\frac{p+1}{p}}$$

In view of Lemma 3, the key ingredient for analyzing the convergence rate in terms of both the objective function gap and the squared gradient norm is a lower bound on $a(t)$. We summarize this result in the following lemma.

Lemma 4. *Suppose that $(x, v, \lambda, a) : [0, +\infty) \mapsto \mathbb{R}^d \times \mathbb{R}^d \times (0, +\infty) \times (0, +\infty)$ is a global solution of the first-order system in Eq. (7). Then, we have*

$$a(t) \geq \left(\frac{c}{2} + \left(\frac{\theta^{\frac{2}{3p+1}}}{(p+1)(\mathcal{E}(0))^{\frac{p-1}{3p+1}}} \right)^{\frac{3p+1}{4}} t^{\frac{3p+1}{4}} \right)^2$$

Proof of Theorem 2. Since the first-order system in Eq. (7) is equivalent to the closed-loop control system in Eq. (3) and Eq. (4), $(x, \lambda, a) : [0, +\infty) \to \mathbb{R}^d \times (0, +\infty) \times (0, +\infty)$ is a global solution of the latter system with $x(0) = x_0 \in \Omega$. By Lemma 3, we have $\mathcal{E}(t) \leq \mathcal{E}(0)$ for $\forall t \geq 0$; that is,

$$a(t)(\Phi(x(t)) - \Phi(x^\star)) + \frac{1}{2}\|v(t) - x^\star\|^2 \leq \mathcal{E}(0)$$

Since $(x(0), v(0)) = (x_0, v_0)$ and $\|v(t) - x^\star\| \geq 0$, we have $a(t)(\Phi(x(t)) - \Phi(x^\star)) \leq \mathcal{E}(0)$. By Lemma 4, we have

$$\Phi(x(t)) - \Phi(x^\star) \leq \mathcal{E}(0) \left(\frac{c}{2} + \left(\frac{\theta^{\frac{2}{3p+1}}}{(p+1)(\mathcal{E}(0))^{\frac{p-1}{3p+1}}} \right)^{\frac{3p+1}{4}} t^{\frac{3p+1}{4}} \right)^{-2} = O(t^{-\frac{3p+1}{2}})$$

By Lemma 3 and using the fact that $\mathcal{E}(t) \geq 0$ for $\forall t \in [0, +\infty)$, we have

$$\int_0^t a(s)\theta^{\frac{1}{p}} \|\nabla \Phi(x(s))\|^{\frac{p+1}{p}} ds \leq \mathcal{E}(0)$$

which implies that

$$\left(\inf_{0\le s\le t}\|\nabla\Phi(x(s))\|^{\frac{p+1}{p}}\right)\left(\int_0^t a(s)ds\right)\le\theta^{-\frac{1}{p}}\mathcal{E}(0)$$

By Lemma 4, we obtain

$$\int_0^t a(s)ds\ge\int_0^t\left(\frac{c}{2}+\left(\frac{\theta^{\frac{2}{3p+1}}}{(p+1)(\mathcal{E}(0))^{\frac{p-1}{3p+1}}}\right)^{\frac{3p+1}{4}}s^{\frac{3p+1}{4}}\right)^2 ds$$

In addition, $\inf_{0\le s\le t}\|\nabla\Phi(x(s))\|^{\frac{p+1}{p}}=(\inf_{0\le s\le t}\|\nabla\Phi(x(s))\|^2)^{\frac{p+1}{2p}}$. Putting these pieces together yields

$$\inf_{0\le s\le t}\|\nabla\Phi(x(s))\|^2\le\left(\frac{\theta^{-\frac{1}{p}}\mathcal{E}(0)}{\int_0^t(\frac{c}{2}+(\frac{\theta^{\frac{2}{3p+1}}}{(p+1)(\mathcal{E}(0))^{\frac{p-1}{3p+1}}})^{\frac{3p+1}{4}}s^{\frac{3p+1}{4}})^2ds}\right)^{\frac{2p}{p+1}}=O(t^{-3p})$$

This completes the proof. □

E Discussion for Section 4

It is useful to compare our approach to approaches based on time scaling [Attouch et al.(2019a), Attouch et al.(2019c), Attouch et al.(2021a), Attouch et al.(2021b)] and quasi-gradient methods [Bégout et al.(201 Attouch et al.(2020a)].

Regularity condition. ***Why is proving the existence and uniqueness of a global solution of the closed-loop control system in Eq.*** (3) ***and Eq.*** (4) ***hard without the regularity condition?*** Our system differs from the existing systems in three respects: (i) the appearance of both $\ddot{x}$ and $\dot{x}$; (ii) the algebraic equation that links λ and $\nabla\Phi(x)$; and (iii) the evolution dynamics depends on λ via a and $\dot{a}$. From a technical point of view, the combination of these features makes it challenging to control a lower bound on gradient norm $\|\nabla\Phi(x(\cdot))\|$ or an upper bound on the feedback control $\lambda(\cdot)$ on the local interval. In sharp contrast, $\|\nabla\Phi(x(t))\|\ge\|\nabla\Phi(x(0))\|e^{-t}$ or $\lambda(t)\le\lambda(0)e^t$ can readily be derived for the Levenberg-Marquardt regularized system in [Attouch and Svaiter(2011), Corollary 3.3] and even the closed-loop control systems without inertia in [Attouch et al.(2013), Theorem 5.2] and [Attouch et al.(2016a), Theorem 2.4]. Thus, we can not exclude the case of $\lambda(t)\to+\infty$ on the bounded interval without the regularity condition and we accordingly fail to establish global existence and uniqueness. We consider it an interesting open problem to derive the regularity condition rather than imposing it as an assumption.

Infinite-dimensional setting. It is promising to study our system using the techniques developed by [Attouch et al.(2016b)] for an infinite-dimensional setting. Our convergence analysis can in fact be extended directly, yielding the same rate of $O(1/t^{(3p+1)/2})$ in terms of objective function gap and $O(1/t^{3p})$ in terms of squared gradient norm in the Hilbert-space setting. However, the weak convergence of the solution trajectories is another matter. Note that [Attouch et al.(2016b)] studied the following open-loop system with the parameters (α,β):

$$\ddot{x}(t)+\frac{\alpha}{t}\dot{x}(t)+\beta\nabla^2\Phi(x(t))\dot{x}(t)+\nabla\Phi(x(t))=0$$

The condition $\alpha > 3$ is crucial for proving weak convergence of solution trajectories and establishing strong convergence in various practical situations. Indeed, the convergence of the solution trajectory has not been established so far when $\alpha = 3$ (except in the one-dimensional case with $\beta = 0$; see [Attouch et al.(2019b)] for the reference). Unfortunately, when $c = 0$ and $p = 1$, the closed-loop control system in Eq. (3) and Eq. (4) becomes

$$\ddot{x}(t) + \frac{3}{t}\dot{x}(t) + \theta\nabla^2\Phi(x(t))\dot{x}(t) + \left(\theta + \frac{\theta}{t}\right)\nabla\Phi(x(t)) = 0$$

The asymptotic damping coefficient $\frac{3}{t}$ does not satisfy the aforementioned condition in [Attouch et al.(2016b)], leaving doubt as to whether weak convergence holds true for the closed-loop control system in Eq. (3) and Eq. (4).

Time scaling. In the context of non-autonomous dissipative systems, time scaling is a simple yet universally powerful tool to accelerate the convergence of solution trajectories [Attouch et al.(2019a), Attouch et al.(2019c), Attouch et al.(2021a), Attouch et al.(2021b)]. Considering the general inertial gradient system in Eq. (3):

$$\ddot{x}(t) + \alpha(t)\dot{x}(t) + \beta(t)\nabla^2\Phi(x(t))\dot{x}(t) + b(t)\nabla\Phi(x(t)) = 0$$

the effect of time scaling is characterized by the coefficient parameter $b(t)$ which comes in as a factor of $\nabla\Phi(x(t))$. In [Attouch et al.(2019a), Attouch et al.(2019c)], the authors conducted an in-depth study of the convergence of this above system without Hessian-driven damping ($\beta = 0$). For the case $\alpha(t) = \frac{\alpha}{t}$, the convergence rate turns out to be $O(\frac{1}{t^2 b(t)})$ under certain conditions on the scalar α and $b(\cdot)$. Thus, a clear improvement can be achieved by taking $b(t) \to +\infty$. This demonstrates the power and potential of time scaling, as further evidenced by recent work on systems with Hessian damping [Attouch et al.(2021a)] and other systems which are associated with the augmented Lagrangian formulation of the affine constrained convex minimization problem [Attouch et al.(2021b)].

Comparing to our closed-loop damping approach, the time scaling technique is based on an open-loop control regime, and indeed $b(t)$ is chosen by hand. In contrast, $\lambda(t)$ in our system is determined by the gradient of $\nabla\Phi(x(t))$ via the algebraic equation, and the evolution dynamics depend on λ via a and $\dot{a}$. The time scaling methodology accordingly does not capture the continuous-time interpretation of optimal acceleration in high-order optimization [Monteiro and Svaiter(2013), Gasnikov et al.(2019), Jiang et al.(2019), Bubeck et al.(2019)]. In contrast, our algebraic equation provides a rigorous justification for the large-step condition in the existing algorithms [Monteiro and Svaiter(2013), Gasnikov et al.(2019), Jiang et al.(2019), Bubeck et al.(2019)] when $p \geq 2$ and demonstrates the fundamental role that the feedback control plays in optimal acceleration, a role clarified by the continuous-time perspective.

Quasi-gradient approach and Kurdyka-Lojasiewicz (KL) theory. The quasi-gradient approach to inertial gradient systems were developed in [Bégout et al.(2015)] and recently applied by [Attouch et al.(2020a)] to analyze inertial dynamics with closed-loop control of the velocity. Recall that a vector field F is called a quasi-gradient for a function E if it has the same singular point as E and if the angle between the field F and the gradient ∇E remains acute and bounded away from $\frac{\pi}{2}$ (see the references [Huang(2006), Chergui(2008), Chill and Fašangová(2010), Bárta et al.(2012), Bárta and Fašangová(2016)] for further geometrical interpretation).

Recent results in [Bégout et al.(2015), Theorem 3.2] and [Attouch et al.(2020a), Theorem 7.2] have suggested that the convergence properties for the bounded trajectories of quasi-gradient systems have been established if the function E is KL [Kurdyka(1998), Bolte et al.(2010)]. In [Attouch et al.(2020a)], the authors considered two closed-loop velocity control systems with a damping potential ϕ:

$$\ddot{x}(t) + \nabla\phi(\dot{x}(t)) + \nabla\Phi(x(t)) = 0. \tag{26}$$

$$\ddot{x}(t) + \nabla\phi(\dot{x}(t)) + \beta\nabla^2\Phi(x(t))\dot{x}(t) + \nabla\Phi(x(t)) = 0. \tag{27}$$

They proposed to use the Hamiltonian formulation of these systems and accordingly defined a function E_λ for $(x, v) = (x, \dot{x}(t))$ by

$$E_\eta(x, v) := \frac{1}{2}\|v\|^2 + \Phi(x) + \eta\langle\nabla\Phi(x), v\rangle$$

If ϕ satisfies some certain growth conditions (see [Attouch et al.(2020a), Theorem 7.3 and 9.2]), the systems in Eq. (26) and Eq. (27) both have a quasi-gradient structure for E_η for sufficiently small $\eta > 0$. This provides an elegant framework for analyzing the convergence properties of the systems in the form of Eq. (26) and Eq. (27) with specific damping potentials.

Why is analyzing our system hard using the quasi-gradient approach? Our system differs from the systems in Eq. (26) and Eq. (27) in two aspects: (i) the closed-loop control law is designed for the gradient of Φ rather than the velocity $\dot{x}$; (ii) the damping coefficients are time dependent, depending on λ via a and $\dot{a}$, and do not have an analytic form when $p \geq 2$. Considering the first-order systems in Eq. (7) and Eq. (8), we find that F is a time-dependent vector field which can not be tackled by the current quasi-gradient approach. We consider it an interesting open problem to develop a quasi-gradient approach for analyzing our system.

F Proofs for Section A

F.1 Proof of Theorem 3

We first present a technical lemma that provides a lower bound for A_k.

Lemma 5. *For $p \geq 1$ and every integer $k \geq 1$, we have*

$$A_k \geq \left(\frac{\theta(1-\sigma^2)^{\frac{p-1}{2}}}{(p+1)^{\frac{3p+1}{2}}\|v_0 - x^\star\|^{p-1}}\right) k^{\frac{3p+1}{2}}$$

Remark. The proof of Lemma 5 is much simpler than the existing analysis; e.g., [Monteiro and Svaiter(2013), Lemma 4.2] for the case of $p = 2$ and [Jiang et al.(2019), Theorem 3.4] and [Bubeck et al.(2019), Lemma 3.3] for the case of $p \geq 2$. Notably, it is not a generalization of the highly technical proof in [Monteiro and Svaiter(2013), Lemma 4.2] but can be interpreted as the discrete-time counterpart of the proof of Lemma 4.

Proof of Theorem 3. For every integer $k \geq 1$, by [Monteiro and Svaiter(2013), Theorem 3.6] and Lemma 5, we have

$$\Phi(x_k) - \Phi(x^\star) \leq \frac{\|v_0 - x^\star\|^2}{2A_k} = O(k^{-\frac{3p+1}{2}})$$

Combining [Monteiro and Svaiter(2013), Proposition 3.9] and Lemma 5, we have

$$\inf_{1\le i\le k} \lambda_i \|w_i\|^2 \le \frac{1+\sigma}{1-\sigma}\frac{\|v_0-x^\star\|^2}{\sum_{i=1}^k A_i} = O(k^{-\frac{3p+3}{2}})$$
$$\inf_{1\le i\le k} \varepsilon_i \le \frac{\sigma^2}{2(1-\sigma^2)}\frac{\|v_0-x^\star\|^2}{\sum_{i=1}^k A_i} = O(k^{-\frac{3p+3}{2}})$$

In addition, we have $\|\lambda_i w_i + x_i - \tilde{v}_{i-1}\| \le \sigma\|x_i - \tilde{v}_{i-1}\|$ and $\lambda_i\|x_i-\tilde{v}_{i-1}\|^{p-1} \ge \theta$. This implies that $\lambda_i\|w_i\|^{\frac{p-1}{p}} \ge \theta^{\frac{1}{p}}(1-\sigma)^{\frac{p-1}{p}}$. Putting these pieces together yields that $\inf_{1\le i\le k}\|w_i\|^{\frac{p+1}{p}} = O(k^{-\frac{3p+3}{2}})$ which implies that

$$\inf_{1\le i\le k}\|w_i\|^2 = \left(\inf_{1\le i\le k}\|w_i\|^{\frac{p+1}{p}}\right)^{\frac{2p}{p+1}} = O(k^{-3p})$$

This completes the proof. □

F.2 Proof of Theorem 4

Inspired by the continuous-time Lyapunov function in Eq. (13), we construct a discrete-time Lyapanunov function for Algorithm 2 as follows:

$$\mathcal{E}_k = \frac{1}{\gamma_k}(\Phi(x_k)-\Phi(x^\star)) + \frac{1}{2}\|v_k - x^\star\|^2 \tag{28}$$

We use this function to prove technical results that pertain to Algorithm 2 and which are the analogs of [Monteiro and Svaiter(2013), Theorem 3.6, Lemma 3.7 and Proposition 3.9].

Lemma 6. *For every integer $k \ge 1$,*

$$\frac{1-\sigma^2}{2}\left(\sum_{i=1}^k \frac{1}{\lambda_i\gamma_i}\|x_i-\tilde{v}_{i-1}\|^2\right) \le \mathcal{E}_0 - \mathcal{E}_k$$

which implies that

$$\Phi(x_k)-\Phi(x^\star) \le \gamma_k\mathcal{E}_0, \quad \|v_k - x^\star\| \le \sqrt{2\mathcal{E}_0}$$

Assuming that $\sigma<1$, we have $\sum_{i=1}^k \frac{1}{\lambda_i\gamma_i}\|x_i-\tilde{v}_{i-1}\|^2 \le \frac{2\mathcal{E}_0}{1-\sigma^2}$.

Lemma 7. *For every integer $k \ge 0$, it holds that*

$$\sqrt{\frac{1}{\gamma_{k+1}}} \ge \sqrt{\frac{1}{\gamma_k}} + \frac{1}{2}\sqrt{\lambda_{k+1}}$$

As a consequence, the following statements hold: (i) For every integer $k \ge 0$, it holds that $\gamma_k \le (1+\frac{1}{2}\sum_{j=1}^k\sqrt{\lambda_j})^{-2}$; (ii) If $\sigma<1$ is further assumed, then we have $\sum_{j=1}^k\|x_j-\tilde{v}_{j-1}\|^2 \le \frac{2\mathcal{E}_0}{1-\sigma^2}$.

Lemma 8. *For every integer $k \ge 1$ and $\sigma<1$, there exists $1\le i\le k$ such that*

$$\inf_{1\le i\le k}\sqrt{\lambda_i}\|w_i\| \le \sqrt{\frac{1+\sigma}{1-\sigma}}\sqrt{\frac{2\mathcal{E}_0}{\sum_{i=1}^k\frac{1}{\gamma_i}}}, \quad \inf_{1\le i\le k}\epsilon_i \le \frac{\sigma^2}{2(1-\sigma^2)}\frac{2\mathcal{E}_0}{\sum_{i=1}^k\frac{1}{\gamma_i}}$$

As the analog of Lemma 5, we provide a technical lemma on the upper bound for γ_k. The analysis is based on the same idea for proving Lemma 5 and is motivated by continuous-time analysis for the first-order system in Eq. (8).

Lemma 9. *For $p \geq 1$ and every integer $k \geq 1$, we have*

$$\gamma_k \leq \frac{(p+1)^{\frac{3p+1}{2}}}{\theta} \left(\frac{2\mathcal{E}_0}{1-\sigma^2}\right)^{\frac{p-1}{2}} k^{-\frac{3p+1}{2}}$$

Proof of Theorem 4. For every integer $k \geq 1$, by Lemma 6 and Lemma 9, we have

$$\Phi(x_k) - \Phi(x^\star) \leq \gamma_k \mathcal{E}_0 = O(k^{-\frac{3p+1}{2}})$$

By Lemma 8 and Lemma 9, we have

$$\inf_{1\leq i\leq k} \lambda_i \|w_i\|^2 \leq \frac{1+\sigma}{1-\sigma} \frac{2\mathcal{E}_0}{\sum_{i=1}^k \frac{1}{\gamma_i}} = O(k^{-\frac{3p+3}{2}})$$
$$\inf_{1\leq i\leq k} \epsilon_i \leq \frac{\sigma^2}{2(1-\sigma^2)} \frac{2\mathcal{E}_0}{\sum_{i=1}^k \gamma_i} = O(k^{-\frac{3p+3}{2}})$$

As in the proof of Theorem 3, we conclude that $\inf_{1\leq i\leq k} \|w_i\|^2 = O(k^{-3p})$. This completes the proof. □

Remark. The discrete-time analysis in this subsection is based on a discrete-time Lyapunov function in Eq. (28), which is closely related to the continuous one in Eq. (13), and two simple yet non-trivial technical lemmas (see Lemma 5 and 9), which are both discrete-time versions of Lemma 4. Notably, the proofs of Lemma 5 and 9 follows the same path for proving Lemma 4 and have demanded the use of the Bihari-LaSalle inequality in discrete time.

F.3 Proof of Theorem 5

Proof of Theorem 5. Given that a pair $(x_k, v_k)_{k\geq 1}$ is generated by Algorithm 3, we define $w_k = \nabla\Phi(x_k)$ and $\varepsilon_k = 0$. Then $v_{k+1} = v_k - a_{k+1}\nabla\Phi(x_{k+1}) = v_k - a_{k+1}w_{k+1}$. Using [Jiang et al.(2019), Proposition 3.2] with a $\hat{\sigma}$-inexact solution $x_{k+1} \in \mathbb{R}^d$ of Eq. (19a) at $(\lambda_{k+1}, \tilde{v}_k)$, a triple $(x_{k+1}, w_{k+1}, \varepsilon_{k+1}) \in \mathbb{R}^d \times \mathbb{R}^d \times (0, +\infty)$ satisfies that

$$w_{k+1} \in \partial_{\epsilon_{k+1}}\Phi(x_{k+1}), \quad \|\lambda_{k+1}w_{k+1} + x_{k+1} - \tilde{v}_k\|^2 + 2\lambda_{k+1}\epsilon_{k+1} \leq \sigma^2\|x_{k+1} - \tilde{v}_k\|^2$$

Since $\theta = \frac{\sigma_l p!}{2\ell} \in (0,1)$ and $\sigma = \hat{\sigma} + \sigma_u < 1$, we have

$$\begin{aligned} \lambda_{k+1}\|x_{k+1} - \tilde{v}_k\|^{p-1} \leq \tfrac{\sigma_u p!}{2\ell} \quad &\Longrightarrow \quad \hat{\sigma} + \tfrac{2\ell\lambda_{k+1}}{p!}\|x_{k+1} - \tilde{v}_k\|^{p-1} \leq \hat{\sigma} + \sigma_u = \sigma \\ \lambda_{k+1}\|x_{k+1} - \tilde{v}_k\|^{p-1} \geq \tfrac{\sigma_l p!}{2\ell} \quad &\Longrightarrow \quad \lambda_{k+1}\|x_{k+1} - \tilde{v}_k\|^{p-1} \geq \theta \end{aligned}$$

Using the same argument with [Bubeck et al.(2019), Lemma 3.1] instead of [Jiang et al.(2019), Proposition 3.2] and an exact solution $x_{k+1} \in \mathbb{R}^d$ of Eq. (19b), we obtain the same result with $\theta = \frac{(p-1)!}{2\ell}$. Putting these pieces together yields the desired conclusion. □

G Proof of Auxiliary Lemmas

G.1 Proof of Lemma 1

Proof of Lemma 1. By the definition, we have

$$\frac{d\mathcal{E}(t)}{dt} = \dot{a}(t)\Phi(x(t)) - \dot{a}(t)\Phi(x^\star) + \langle a(t)\dot{x}(t), \nabla\Phi(x(t))\rangle + \langle \dot{v}(t), v(t) - x^\star\rangle$$

In addition, we have $\langle \dot{v}(t), v(t)-x^\star\rangle = \langle \dot{v}(t), v(t)-x(t)\rangle + \langle \dot{v}(t), x(t)-x^\star\rangle$ and $\dot{v}(t) = -\dot{a}(t)\nabla\Phi(x(t))$. Putting these pieces together yields:

$$\begin{aligned}\frac{d\mathcal{E}(t)}{dt} &= \underbrace{\dot{a}(t)(\Phi(x(t)) - \Phi(x^\star) - \langle \nabla\Phi(x(t)), x(t) - x^\star\rangle)}_{\mathbf{I}} \\ &+ \underbrace{\langle a(t)\dot{x}(t), \nabla\Phi(x(t))\rangle + \dot{a}(t)\langle x(t) - v(t), \nabla\Phi(x(t))\rangle}_{\mathbf{II}}\end{aligned}$$

By the convexity of Φ, we have $\Phi(x(t)) - \Phi(x^\star) - \langle \nabla\Phi(x(t)), x(t) - x^\star\rangle \le 0$. Since $\dot{a}(t) \ge 0$, we have $\mathbf{I} \le 0$. Furthermore, Eq. (7) implies that

$$\dot{x}(t) + \frac{\dot{a}(t)}{a(t)}(x(t) - v(t)) = -\lambda(t)\nabla\Phi(x(t))$$

which implies that

$$\mathbf{II} = \langle a(t)\dot{x}(t) + \dot{a}(t)x(t) - \dot{a}(t)v(t), \nabla\Phi(x(t))\rangle = -\lambda(t)a(t)\|\nabla\Phi(x(t))\|^2$$

This together with the algebraic equation implies $\mathbf{II} \le -a(t)\theta^{\frac{1}{p}}\|\nabla\Phi(x(t))\|^{\frac{p+1}{p}}$. Putting all these pieces together yields the desired inequality. □

G.2 Proof of Lemma 2

Proof of Lemma 2. By Lemma 1, the function $\mathcal{E}$ is nonnegative and nonincreasing on the interval $[0, t_0]$. This implies that, for any $t \in [0, t_0]$, we have

$$\frac{1}{2}\|v(t) - x^\star\|^2 \le a(t)(\Phi(x(t)) - \Phi(x^\star)) + \frac{1}{2}\|v(t) - x^\star\|^2 \le \mathcal{E}(0)$$

Therefore, $v(\cdot)$ is bounded on the interval $[0, t_0]$ and the upper bound only depends on the initial condition. Furthermore, we have

$$a(t)(x(t) - x^\star) - a(0)(x_0 - x^\star) = \int_0^t (\dot{a}(s)(x(s) - x^\star) + a(s)\dot{x}(s))ds$$

Using the triangle inequality and $a(0) = c^2$, we have

$$\begin{aligned}\|a(t)(x(t) - x^\star)\| &\le c^2\|x_0 - x^\star\| + \int_0^t \|a(s)\dot{x}(s) + \dot{a}(t)x(s) - \dot{a}(s)x^\star\|ds \\ &\overset{\text{Eq. (7)}}{\le} c^2\|x_0 - x^\star\| + \int_0^t \|\dot{a}(s)v(s) - \dot{a}(s)x^\star\|ds + \int_0^t \|\lambda(s)a(s)\nabla\Phi(x(s))\|ds\end{aligned}$$

Note that $\|v(t)-x^\star\| \le \sqrt{2\mathcal{E}(0)}$ is proved for all $t \in [0,t_0]$ and $a(t)$ is monotonically increasing with $a(0)=c^2$. Thus, the following inequality holds:

$$\begin{aligned}\|x(t)-x^\star\| &\le \frac{c^2\|x_0-x^\star\| + (a(t)-c^2)\sqrt{2\mathcal{E}(0)} + \int_0^t \lambda(s)a(s)\|\nabla\Phi(x(s))\|ds}{a(t)} \\ &\le \|x_0-x^\star\| + \sqrt{2\mathcal{E}(0)} + \frac{1}{a(t)}\int_0^t \lambda(s)a(s)\|\nabla\Phi(x(s))\|ds\end{aligned}$$

By the Hölder inequality and using the fact that $a(t)$ is monotonically increasing, we have

$$\begin{aligned}&\int_0^t \lambda(s)a(s)\|\nabla\Phi(x(s))\|ds = \int_0^t \sqrt{\lambda(s)a(s)}(\sqrt{\lambda(s)a(s)}\|\nabla\Phi(x(s))\|)ds \\ &\le \left(\int_0^t \lambda(s)a(s)ds\right)^{1/2}\left(\int_0^t \lambda(s)a(s)\|\nabla\Phi(x(s))\|^2ds\right)^{1/2} \\ &\le \sqrt{a(t)}\left(\int_0^t \sqrt{\lambda(s)}ds\right)\left(\int_0^t \lambda(s)a(s)\|\nabla\Phi(x(s))\|^2ds\right)^{1/2} \\ &\le a(t)\left(\int_0^t \lambda(s)a(s)\|\nabla\Phi(x(s))\|^2ds\right)^{1/2}\end{aligned}$$

The algebra equation implies that $\lambda(t)\|\nabla\Phi(x(t))\|^2 = \theta^{\frac{1}{p}}\|\nabla\Phi(x(t))\|^{\frac{p+1}{p}}$. Thus, by Lemma 1 again, we have

$$\int_0^t \lambda(s)a(s)\|\nabla\Phi(x(s))\|^2ds = \int_0^t a(s)\theta^{\frac{1}{p}}\|\nabla\Phi(x(s))\|^{\frac{p+1}{p}}ds \le \mathcal{E}(0)$$

Putting these pieces together yields that $\|x(t)-x^\star\| \le \|x_0-x^\star\| + 3\sqrt{\mathcal{E}(0)}$. Therefore, $x(t)$ is bounded on the interval $[0,t_0]$ and the upper bound only depends on the initial condition. This completes the proof. □

G.3 Proof of Lemma 4

Proof of Lemma 4. For $p=1$, the feedback control law is given by $\lambda(t)=\theta$, for $\forall t\in[0,+\infty)$, and

$$a(t) = \left(\frac{c}{2}+\frac{\sqrt{\theta}t}{2}\right)^2 = \left(\frac{c}{2}+\left(\frac{\theta^{\frac{2}{3p+1}}}{(p+1)(\mathcal{E}(0))^{\frac{p-1}{3p+1}}}\right)^{\frac{3p+1}{4}} t^{\frac{3p+1}{4}}\right)^2$$

For $p\ge 2$, the algebraic equation implies that $\|\nabla\Phi(x(t))\| = (\frac{\theta^{1/p}}{\lambda(t)})^{\frac{p}{p-1}}$ since $\lambda(t)>0$ for $\forall t\in[0,+\infty)$. This together with Lemma 3 implies that

$$\frac{d\mathcal{E}(t)}{dt} \le -a(t)\theta^{\frac{1}{p}}\|\nabla\Phi(x(t))\|^{\frac{p+1}{p}} = -a(t)\theta^{\frac{2}{p-1}}[\lambda(t)]^{-\frac{p+1}{p-1}}$$

Since $\mathcal{E}(t)\ge 0$, we have

$$\int_0^t a(s)\theta^{\frac{2}{p-1}}(\lambda(s))^{-\frac{p+1}{p-1}}ds \le \mathcal{E}(0)$$

By the Hölder inequality, we have

$$\int_0^t (a(s))^{\frac{p-1}{3p+1}} ds = \int_0^t (a(s)(\lambda(s))^{-\frac{p+1}{p-1}})^{\frac{p-1}{3p+1}} (\lambda(s))^{\frac{p+1}{3p+1}} ds$$
$$\leq \left(\int_0^t a(s)(\lambda(s))^{-\frac{p+1}{p-1}} ds \right)^{\frac{p-1}{3p+1}} \left(\int_0^t \sqrt{\lambda(s)} ds \right)^{\frac{2p+2}{3p+1}}$$

Combining these results with the definition of a yields:

$$\int_0^t (a(s))^{\frac{p-1}{3p+1}} ds \leq \theta^{-\frac{2}{3p+1}} (\mathcal{E}(0))^{\frac{p-1}{3p+1}} \left(\int_0^t \sqrt{\lambda(s)} ds \right)^{\frac{2p+2}{3p+1}}$$
$$\leq \theta^{-\frac{2}{3p+1}} (\mathcal{E}(0))^{\frac{p-1}{3p+1}} (2\sqrt{a(t)} - c)^{\frac{2p+2}{3p+1}} \leq 2\theta^{-\frac{2}{3p+1}} (\mathcal{E}(0))^{\frac{p-1}{3p+1}} \left(\sqrt{a(t)} - \frac{c}{2} \right)^{\frac{2p+2}{3p+1}}$$

Since $a(t)$ is nonnegative and nondecreasing with $\sqrt{a(0)} = \frac{c}{2}$, we have

$$\int_0^t \left(\sqrt{a(s)} - \frac{c}{2} \right)^{\frac{2p-2}{3p+1}} ds \leq 2\theta^{-\frac{2}{3p+1}} (\mathcal{E}(0))^{\frac{p-1}{3p+1}} \left(\sqrt{a(t)} - \frac{c}{2} \right)^{\frac{2p+2}{3p+1}} \tag{29}$$

The remaining steps in the proof are based on the Bihari-LaSalle inequality [LaSalle(1949), Bihari(1956)]. In particular, we denote $y(\cdot)$ by $y(t) = \int_0^t (\sqrt{a(s)} - \frac{c}{2})^{\frac{2p-2}{3p+1}} ds$. Then, $y(0) = 0$ and Eq. (29) implies that

$$y(t) \leq 2\theta^{-\frac{2}{3p+1}} (\mathcal{E}(0))^{\frac{p-1}{3p+1}} (\dot{y}(t))^{\frac{p+1}{p-1}}$$

This implies that

$$\dot{y}(t) \geq \left(\frac{y(t)}{2\theta^{-\frac{2}{3p+1}} (\mathcal{E}(0))^{\frac{p-1}{3p+1}}} \right)^{\frac{p-1}{p+1}} \Longrightarrow \frac{\dot{y}(t)}{(y(t))^{\frac{p-1}{p+1}}} \geq \left(\frac{1}{2\theta^{-\frac{2}{3p+1}} (\mathcal{E}(0))^{\frac{p-1}{3p+1}}} \right)^{\frac{p-1}{p+1}}$$

Integrating this inequality over $[0, t]$ yields:

$$(y(t))^{\frac{2}{p+1}} \geq \frac{2}{p+1} \left(\frac{1}{2\theta^{-\frac{2}{3p+1}} (\mathcal{E}(0))^{\frac{p-1}{3p+1}}} \right)^{\frac{p-1}{p+1}} t$$

Equivalently, by the definition of $y(t)$, we have

$$\int_0^t \left(\sqrt{a(s)} - \frac{c}{2} \right)^{\frac{2p-2}{3p+1}} ds \geq \left(\frac{2}{p+1} \right)^{\frac{p+1}{2}} \left(\frac{1}{2\theta^{-\frac{2}{3p+1}} (\mathcal{E}(0))^{\frac{p-1}{3p+1}}} \right)^{\frac{p-1}{2}} t^{\frac{p+1}{2}}$$

This together with Eq. (29) yields that

$$\sqrt{a(t)} \geq \frac{c}{2} + \left(\frac{1}{2\theta^{-\frac{2}{3p+1}} (\mathcal{E}(0))^{\frac{p-1}{3p+1}}} \int_0^t \left(\sqrt{a(s)} - \frac{c}{2} \right)^{\frac{2p-2}{3p+1}} ds \right)^{\frac{3p+1}{2p+2}}$$
$$\geq \frac{c}{2} + \left(\frac{\theta^{\frac{2}{3p+1}}}{(p+1)(\mathcal{E}(0))^{\frac{p-1}{3p+1}}} \right)^{\frac{3p+1}{4}} t^{\frac{3p+1}{4}}$$

This completes the proof. □

G.4 Proof of Lemma 5

Proof of Lemma 5. For $p = 1$, the large-step condition implies that $\lambda_k \geq \theta$ for all $k \geq 0$. By [Monteiro and Svaiter(2
Lemma 3.7], we have $A_k \geq \frac{\theta k^2}{4}$.

For $p \geq 2$, the large-step condition implies that

$$\sum_{i=1}^{k} A_i(\lambda_i)^{-\frac{p+1}{p-1}}\theta^{\frac{2}{p-1}} \leq \sum_{i=1}^{k} A_i(\lambda_i)^{-\frac{p+1}{p-1}}(\lambda_i\|x_i - \tilde{v}_{i-1}\|^{p-1})^{\frac{2}{p-1}}$$
$$= \sum_{i=1}^{k} \frac{A_i}{\lambda_i}\|x_i - \tilde{v}_{i-1}\|^2 \overset{\text{[Monteiro and Svaiter(2013), Theorem 3.6]}}{\leq} \frac{\|v_0 - x^\star\|^2}{1-\sigma^2}$$

By the Hölder inequality, we have

$$\sum_{i=1}^{k}(A_i)^{\frac{p-1}{3p+1}} = \sum_{i=1}^{k}(A_i(\lambda_i)^{-\frac{p+1}{p-1}})^{\frac{p-1}{3p+1}}(\lambda_i)^{\frac{p+1}{3p+1}} \leq (\sum_{i=1}^{k} A_i(\lambda_i)^{-\frac{p+1}{p-1}})^{\frac{p-1}{3p+1}}(\sum_{i=1}^{k}\sqrt{\lambda_i})^{\frac{2p+2}{3p+1}}$$

For the ease of presentation, we define $C = \theta^{-\frac{2}{3p+1}}(\frac{\|v_0-x^\star\|^2}{1-\sigma^2})^{\frac{p-1}{3p+1}}$. Putting these pieces together yields:

$$\sum_{i=1}^{k}(A_i)^{\frac{p-1}{3p+1}} \leq C\left(\sum_{i=1}^{k}\sqrt{\lambda_i}\right)^{\frac{2p+2}{3p+1}} \overset{\text{[Monteiro and Svaiter(2013), Lemma 3.7]}}{\leq} 2C(A_k)^{\frac{p+1}{3p+1}} \tag{30}$$

The remaining proof is based on the Bihari-LaSalle inequality in discrete time. In particular, we define $\{y_k\}_{k\geq 0}$ by $y_k = \sum_{i=1}^{k}(A_i)^{\frac{p-1}{3p+1}}$. Then, $y_0 = 0$ and Eq. (30) implies that

$$y_k \leq 2C(y_k - y_{k-1})^{\frac{p+1}{p-1}}$$

This implies that

$$y_k - y_{k-1} \geq \left(\frac{y_k}{2C}\right)^{\frac{p-1}{p+1}} \Longrightarrow \frac{y_k - y_{k-1}}{(y_k)^{\frac{p-1}{p+1}}} \geq \left(\frac{1}{2C}\right)^{\frac{p-1}{p+1}} \tag{31}$$

Inspired by the continuous-time inequality in Lemma 5, we claim that the following discrete-time inequality holds for every integer $k \geq 1$:

$$(y_k)^{\frac{2}{p+1}} - (y_{k-1})^{\frac{2}{p+1}} \geq \frac{2}{p+1}\left(\frac{y_k - y_{k-1}}{[y_k]^{\frac{p-1}{p+1}}}\right) \tag{32}$$

Indeed, we define $g(t) = 1 - t^{\frac{2}{p+1}}$ and find that this function is convex for $\forall t \in (0,1)$ since $p \geq 1$. Thus, we have

$$1 - t^{\frac{2}{p+1}} = g(t) - g(1) \geq (t-1)\nabla g(1) = \frac{2(1-t)}{p+1} \Longrightarrow \frac{1-t^{\frac{2}{p+1}}}{1-t} \geq \frac{2}{p+1}$$

Since y_k is increasing, we have $\frac{y_{k-1}}{y_k} \in (0,1)$. Then, the desired Eq. (31) follows from setting $t = \frac{y_{k-1}}{y_k}$. Combining Eq. (31) and Eq. (32) yields that

$$(y_k)^{\frac{2}{p+1}} - (y_{k-1})^{\frac{2}{p+1}} \geq \frac{2}{p+1}\left(\frac{1}{2C}\right)^{\frac{p-1}{p+1}}$$

Therefore, we conclude that

$$(y_k)^{\frac{2}{p+1}} = (y_0)^{\frac{2}{p+1}} + \left(\sum_{i=1}^{k}(y_i)^{\frac{2}{p+1}} - (y_{i-1})^{\frac{2}{p+1}}\right) \geq \frac{2}{p+1}\left(\frac{1}{2C}\right)^{\frac{p-1}{p+1}} k$$

By the definition of y_k, we have

$$\sum_{i=1}^{k}(A_i)^{\frac{p-1}{3p+1}} \geq \left(\frac{2}{p+1}\right)^{\frac{p+1}{2}}\left(\frac{1}{2C}\right)^{\frac{p-1}{2}} k^{\frac{p+1}{2}}$$

This together with Eq. (30) yields that

$$A_k \geq \left(\frac{1}{2C}\sum_{i=1}^{k}(A_i)^{\frac{p-1}{3p+1}}\right)^{\frac{3p+1}{p+1}} \geq \left(\frac{1}{(p+1)C}\right)^{\frac{3p+1}{2}} k^{\frac{3p+1}{2}}$$

This completes the proof. □

G.5 Proof of Lemma 6

Proof of Lemma 6. It suffices to prove the first inequality which implies the other results. Based on the discrete-time Lyapunov function, we define two functions $\phi_k : \mathbb{R}^d \mapsto \mathbb{R}$ and $\Gamma_k : \mathbb{R}^d \mapsto \mathbb{R}$ by (Γ_k is related to $\mathcal{E}_k$ and defined recursively):

$$\phi_k(v) = \Phi(x_k) + \langle v - x_k, w_k\rangle - \epsilon_k - \Phi(x^\star),\ \forall k \geq 0$$
$$\Gamma_0(v) = \frac{1}{\gamma_0}(\Phi(x_0) - \Phi(x^\star)) + \frac{1}{2}\|v - v_0\|^2,\ \Gamma_{k+1} = \Gamma_k + \frac{\alpha_{k+1}}{\gamma_{k+1}}\phi_{k+1},\ \forall k \geq 0$$

First, by definition, ϕ_k is affine. Since $w_{k+1} \in \partial_{\epsilon_{k+1}}\Phi(x_{k+1})$, Eq. (15) implies that $\phi_k(v) \leq \Phi(v) - \Phi(x^\star)$. Furthermore, Γ_k is quadratic and $\nabla^2\Gamma_k = \nabla^2\Gamma_0$ since ϕ_k is affine. Then, we prove that $\Gamma_k(v) \leq \Gamma_0(v) + \frac{1-\gamma_k}{\gamma_k}(\Phi(v) - \Phi(x^\star))$ using induction. Indeed, it holds when $k = 0$ since $\gamma_0 = 1$. Assuming that this inequality holds for $\forall i \leq k$, we derive from $\phi_k(v) \leq \Phi(v) - \Phi(x^\star)$ and $\gamma_{k+1} = (1 - \alpha_{k+1})\gamma_k$ that

$$\Gamma_{k+1}(v) \leq \Gamma_0(v) + \left(\frac{1-\gamma_k}{\gamma_k} + \frac{\alpha_{k+1}}{\gamma_{k+1}}\right)(\Phi(v) - \Phi(x^\star)) = \Gamma_0(v) + \frac{1-\gamma_k}{\gamma_k}(\Phi(v) - \Phi(x^\star))$$

Finally, we prove that $v_k = \operatorname{argmin}_{v\in\mathbb{R}^d}\Gamma_k(v)$ using the induction. Indeed, it holds when $k = 0$. Suppose that this inequality holds for $\forall i \leq k$, we have

$$\nabla\Gamma_{k+1}(v) = \nabla\Gamma_k(v) + \frac{\alpha_{k+1}}{\gamma_{k+1}}\nabla\phi_{k+1}(v) = v - v_k + \frac{\alpha_{k+1}}{\gamma_{k+1}}w_{k+1}$$

Using the definition of v_k and the fact that $\gamma_{k+1} = (1-\alpha_{k+1})\gamma_k$, we have $\nabla\Gamma_{k+1}(v) = 0$ if and only if $v = v_{k+1}$.

The remaining proof is based on the gap sequence $\{\beta_k\}_{k\geq 0}$ which is defined by $\beta_k = \inf_{v\in\mathbb{R}^d}\Gamma_k(v) - \frac{1}{\gamma_k}(\Phi(x_k) - \Phi(x^\star))$. Using the previous facts that Γ_k is quadratic with $\nabla^2\Gamma_k = 1$ and the upper bound for $\Gamma_k(v)$, we have

$$\beta_k = \Gamma_k(x^\star) - \frac{1}{\gamma_k}(\Phi(x_k) - \Phi(x^\star)) - \frac{1}{2}\|x^\star - v_k\|^2 \leq \Gamma_0(x^\star) - \mathcal{E}_k = \mathcal{E}_0 - \mathcal{E}_k$$

By definition, we have $\beta_0 = 0$. Thus, it suffices to prove that the following recursive inequality holds true for every integer $k \geq 0$,

$$\beta_{k+1} \geq \beta_k + \frac{1-\sigma^2}{2\lambda_{k+1}\gamma_{k+1}}\|x_{k+1} - \tilde{v}_k\|^2 \tag{33}$$

In particular, we define $\tilde{v} = (1-\alpha_{k+1})x_k + \alpha_{k+1}v$ for any given $v \in \mathbb{R}^d$. Using the definition of $\tilde{v}_k$ and the affinity of ϕ_{k+1}, we have

$$\phi_{k+1}(\tilde{v}) = (1-\alpha_{k+1})\phi_{k+1}(x_k) + \alpha_{k+1}\phi_{k+1}(v), \tag{34}$$

$$\tilde{v} - \tilde{v}_k = \alpha_{k+1}(v - v_k). \tag{35}$$

Since Γ_k is quadratic with $\nabla^2\Gamma_k = 1$, we have $\Gamma_k(v) = \Gamma_k(v_k) + \frac{1}{2}\|v - v_k\|^2$. Plugging this into the recursive equation for Γ_k yields that

$$\Gamma_{k+1}(v) = \Gamma_k(v_k) + \frac{1}{2}\|v - v_k\|^2 + \frac{\alpha_{k+1}}{\gamma_{k+1}}\phi_{k+1}(v)$$

By the definition of β_k, we have $\Gamma_k(v_k) = \beta_k + \frac{1}{\gamma_k}(\Phi(x_k) - \Phi(x^\star))$. Putting these pieces together with the definition of $\mathcal{E}_k$ yields that

$$\Gamma_{k+1}(v) = \beta_k + \frac{\alpha_{k+1}}{\gamma_{k+1}}\phi_{k+1}(v) + \frac{1}{\gamma_k}(\Phi(x_k) - \Phi(x^\star)) + \frac{1}{2}\|v - v_k\|^2$$

Since $\phi_{k+1}(v) \leq \Phi(v) - \Phi(x^\star)$, we have

$$\begin{aligned}
\Gamma_{k+1}(v) &\geq \beta_k + \frac{\alpha_{k+1}}{\gamma_{k+1}}\phi_{k+1}(v) + \frac{1}{\gamma_k}\phi_{k+1}(x_k) + \frac{1}{2}\|v - v_k\|^2 \\
&\overset{\text{Eq. (34)}}{=} \beta_k + \frac{1}{\gamma_{k+1}}\phi_{k+1}(\tilde{v}) + \frac{1}{2}\|v - v_k\|^2 \\
&= \beta_k + \frac{1}{\gamma_{k+1}}\left(\phi_{k+1}(\tilde{v}) + \frac{\gamma_{k+1}}{2}\|v - v_k\|^2\right) \\
&\overset{\text{Eq. (35)}}{=} \beta_k + \frac{1}{\gamma_{k+1}}\left(\phi_{k+1}(\tilde{v}) + \frac{\gamma_{k+1}}{2(\alpha_{k+1})^2}\|\tilde{v} - \tilde{v}_k\|^2\right) \\
&= \beta_k + \frac{1}{\gamma_{k+1}}\left(\phi_{k+1}(\tilde{v}) + \frac{1}{2\lambda_{k+1}}\|\tilde{v} - \tilde{v}_k\|^2\right)
\end{aligned}$$

Using [Monteiro and Svaiter(2013), Lemma 3.3] with $\lambda = \lambda_{k+1}$, $\tilde{v} = \tilde{v}_k$, $\tilde{x} = x_{k+1}$, $\tilde{w} = w_{k+1}$ and $\epsilon = \epsilon_{k+1}$, we have

$$\inf_{v\in\mathbb{R}^d}\left\{\langle v - x_{k+1}, w_{k+1}\rangle - \epsilon_{k+1} + \frac{1}{2\lambda_{k+1}}\|v - \tilde{v}_k\|^2\right\} \geq \frac{1-\sigma^2}{2\lambda_{k+1}}\|x_{k+1} - \tilde{v}_k\|^2$$

which implies that

$$\phi_{k+1}(\tilde{v}) + \frac{1}{2\lambda_{k+1}}\|\tilde{v} - \tilde{v}_k\|^2 - \frac{1}{\gamma_{k+1}}(\Phi(x_{k+1}) - \Phi(x^\star)) \geq \frac{1-\sigma^2}{2\lambda_{k+1}}\|x_{k+1} - \tilde{v}_k\|^2$$

Putting these pieces together yields that

$$\inf_{v\in\mathbb{R}^d}\Gamma_{k+1}(v) - \frac{1}{\gamma_{k+1}}(\Phi(x_{k+1}) - \Phi(x^\star)) \geq \beta_k + \frac{1-\sigma^2}{2\lambda_{k+1}\gamma_{k+1}}\|x_{k+1} - \tilde{v}_k\|^2$$

which together with the definition of β_k yields the desired inequality in Eq. (33). This completes the proof. □

G.6 Proof of Lemma 7

Proof of Lemma 7. It suffices to prove the first inequality which implies the other results. By the definition of $\{\gamma_k\}_{k\geq 0}$ and $\{\alpha_k\}_{k\geq 0}$, we have $\gamma_{k+1} = (1-\alpha_{k+1})\gamma_k$ and $(\alpha_{k+1})^2 = \lambda_{k+1}\gamma_{k+1}$. This implies that

$$\frac{1}{\gamma_k} = \frac{1}{\gamma_{k+1}} - \frac{\alpha_{k+1}}{\gamma_{k+1}} = \frac{1}{\gamma_{k+1}} - \sqrt{\frac{\lambda_{k+1}}{\gamma_{k+1}}}$$

Since $\gamma_k > 0$ and $\lambda_k > 0$, we have $\sqrt{\frac{1}{\gamma_{k+1}}} \geq \frac{1}{2}\sqrt{\lambda_{k+1}}$ and

$$\frac{1}{\gamma_k} \leq \frac{1}{\gamma_{k+1}} - \sqrt{\frac{\lambda_{k+1}}{\gamma_{k+1}}} + \frac{\lambda_{k+1}}{4} = \left(\sqrt{\frac{1}{\gamma_{k+1}}} - \frac{1}{2}\sqrt{\lambda_{k+1}}\right)^2$$

which implies the desired inequality. □

G.7 Proof of Lemma 8

Proof of Lemma 8. With the convention $0/0 = 0$, we define $\tau_k = \max\{\frac{2\epsilon_k}{\sigma^2}, \frac{\lambda_k\|w_k\|^2}{(1+\sigma)^2}\}$ for every integer $k \geq 1$. Then, we have

$$\begin{aligned}
&2\lambda_k\epsilon_k \leq \sigma^2\|x_k - \tilde{v}_{k-1}\|^2 \\
&\|\lambda_k w_k\| \leq \|\lambda_k w_k + x_k - \tilde{v}_{k-1}\| + \|x_k - \tilde{v}_{k-1}\| \leq (1+\sigma)\|x_k - \tilde{v}_{k-1}\|
\end{aligned}$$

which implies that $\lambda_k\tau_k \leq \|x_k - \tilde{v}_{k-1}\|^2$ for every integer $k \geq 1$. This together with Lemma 6 yields that

$$\frac{2\mathcal{E}_0}{1-\sigma^2} \geq \sum_{i=1}^{k} \frac{1}{\lambda_i\gamma_i}\|x_i - \tilde{v}_{i-1}\|^2 \geq \left(\inf_{1\leq i\leq k} \tau_i\right)\left(\sum_{i=1}^{k}\frac{1}{\gamma_i}\right)$$

Combining this inequality with the definition of τ_k yields the desired results. □

G.8 Proof of Lemma 9

Proof of Lemma 9. For $p = 1$, the large-step condition implies that $\lambda_k \geq \theta$ for all $k \geq 0$. By Lemma 7, we have $\gamma_k \leq \frac{4}{\theta k^2}$.

For $p \geq 2$, the large-step condition implies that

$$\begin{aligned}
\sum_{i=1}^{k}(\gamma_i)^{-1}(\lambda_i)^{-\frac{p+1}{p-1}}\theta^{\frac{2}{p-1}} &\leq \sum_{i=1}^{k}(\gamma_i)^{-1}(\lambda_i)^{-\frac{p+1}{p-1}}(\lambda_i\|x_i - \tilde{v}_{i-1}\|^{p-1})^{\frac{2}{p-1}} \\
&= \sum_{i=1}^{k}\frac{1}{\lambda_i\gamma_i}\|x_i - \tilde{v}_{i-1}\|^2 \overset{\text{Lemma 6}}{\leq} \frac{2\mathcal{E}_0}{1-\sigma^2}
\end{aligned}$$

By the Hölder inequality, we have

$$\sum_{i=1}^{k}(\gamma_i)^{-\frac{p-1}{3p+1}} = \sum_{i=1}^{k}\left(\frac{1}{(\lambda_i)^{\frac{p+1}{p-1}}\gamma_i}\right)^{\frac{p-1}{3p+1}}(\lambda_i)^{\frac{p+1}{3p+1}} \leq \left(\sum_{i=1}^{k}\frac{1}{(\lambda_i)^{\frac{p+1}{p-1}}\gamma_i}\right)^{\frac{p-1}{3p+1}}\left(\sum_{i=1}^{k}\sqrt{\lambda_i}\right)^{\frac{2p+2}{3p+1}}$$

For ease of presentation, we define $C = \theta^{-\frac{2}{3p+1}} (\frac{2\mathcal{E}_0}{1-\sigma^2})^{\frac{p-1}{3p+1}}$. Putting these pieces together yields that

$$\sum_{i=1}^{k} (\gamma_i)^{-\frac{p-1}{3p+1}} \leq C \left(\sum_{i=1}^{k} \sqrt{\lambda_i} \right)^{\frac{2p+2}{3p+1}} \overset{\text{Lemma 7}}{\leq} 2C(\gamma_k)^{-\frac{p+1}{3p+1}} \tag{36}$$

Using the same argument for proving Lemma 5, we have

$$\sum_{i=1}^{k} (\gamma_i)^{-\frac{p-1}{3p+1}} \geq \left(\frac{2}{p+1} \right)^{\frac{p+1}{2}} \left(\frac{1}{2C} \right)^{\frac{p-1}{2}} k^{\frac{p+1}{2}}$$

This together with Eq. (36) yields that

$$\frac{1}{\gamma_k} \geq \left(\frac{1}{2C} \sum_{i=1}^{k} (\gamma_i)^{-\frac{p-1}{3p+1}} \right)^{\frac{3p+1}{p+1}} \geq \left(\frac{1}{(p+1)C} \right)^{\frac{3p+1}{2}} k^{\frac{3p+1}{2}}$$

This completes the proof. □