Mathematical Programming (2022) 195:929–975
https://doi.org/10.1007/s10107-021-01721-3

FULL LENGTH PAPER
Series A

# A control-theoretic perspective on optimal high-order optimization

Tianyi Lin[1] · Michael I. Jordan[1,2]

Received: 21 December 2019 / Accepted: 4 October 2021 / Published online: 22 October 2021

## Abstract

We provide a control-theoretic perspective on optimal tensor algorithms for minimizing a convex function in a finite-dimensional Euclidean space. Given a function $\Phi : \mathbb{R}^d \to \mathbb{R}$ that is convex and twice continuously differentiable, we study a closed-loop control system that is governed by the operators $\nabla\Phi$ and $\nabla^2\Phi$ together with 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 first contribution is to prove the existence and uniqueness of a local solution to this system via the Banach fixed-point theorem. We present a simple yet nontrivial Lyapunov function that allows us to establish the existence and uniqueness of a global solution under certain regularity conditions and analyze the convergence properties of trajectories. The rate of convergence is $O(1/t^{(3p+1)/2})$ in terms of objective function gap and $O(1/t^{3p})$ in terms of squared gradient norm. Our second contribution is to provide two algorithmic frameworks obtained from discretization of our continuous-time system, one of which generalizes the large-step A-HPE framework of Monteiro and Svaiter (SIAM J Optim 23(2):1092–1125, 2013) and the other of which leads to a new optimal $p$-th order tensor algorithm. While our discrete-time analysis can be seen as a simplification and generalization of Monteiro and Svaiter (2013), it is largely motivated by the aforementioned continuous-time analysis, demonstrating the fundamental role that the feedback control plays in optimal acceleration and the clear advantage that the continuous-time perspective brings to algorithmic design. A highlight of our analysis is that we show that all of the $p$-th order optimal tensor algorithms that we discuss minimize the squared gradient norm at a rate of $O(k^{-3p})$, which complements the recent analysis in Gasnikov et al. (in: COLT, PMLR, pp 1374–1391, 2019), Jiang et al. (in: COLT, PMLR, pp 1799–1801, 2019) and Bubeck et al. (in: COLT, PMLR, pp 492–507, 2019).

✉ Tianyi Lin
darren_lin@cs.berkeley.edu

Michael I. Jordan
jordan@cs.berkeley.edu

[1] Department of Electrical Engineering and Computer Science, UC Berkeley, Berkeley, USA

[2] Department of Statistics, UC Berkeley, Berkeley, USA

**Keywords** Convex optimization · Optimal acceleration · Closed-loop control system · Feedback control · High-order tensor algorithm · Iteration complexity

**Mathematics Subject Classification** 37N40 · 90C25 · 90C60 · 49M37 · 68Q25

## 1 Introduction

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 [4,5,10,24,25,98]; and (2) the explicit discretization of Newton-like and Levenberg–Marquardt regularized systems [1,6,7,12,26–28,32–34,79], 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) [95] and related methods [38,67,90,108]. These perspectives arise from derivations that obtain differential equations as limits of discrete dynamics [29,30,56,74,86,101,102,106,109], including quasi-gradient formulations and Kurdyka-Lojasiewicz theory [14,39] (see [36,37,52,53,69] for geometrical perspective on the topic), inertial gradient systems with constant or asymptotic vanishing damping [15,20,21,106] and their extension to maximally monotone operators [16,17,45], Hessian-driven damping [6,13,18,28,31,46,102], time scaling [13,19,21,22], dry friction damping [2,3], closed-loop damping [13,14], control-theoretic design [58,68,77] and Lagrangian and Hamiltonian frameworks [40,55,59,60,78,87,96,110]. 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 [20,29,83] and $O(k^{-3})$ in terms of squared gradient norm [102].

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 [7], a development that inspired work on continuous-time Newton-like approaches for convex minimization [7,32] and monotone inclusions [1,12,26,27,33,34,79]. Building on these works, [6] distinguished Hessian-driven damping from classical continuous Newton formulations and showed its importance in optimization and mechanics. Subsequently, [31] 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 [106]. 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 [102], 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 Eqs. (1) and (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 [13,14]. 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 [11,47,61,71,85,91,105], 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 [12,33], 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 $(\alpha, \beta, b)$ explicitly depends on the variables $(x, \lambda, a)$, the parameters $c > 0$, $\theta \in (0, 1)$ and the order $p \in \{1, 2, \ldots\}$:

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

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 [12] for modeling the proximal Newton algorithm. We recall that Eq. (3) has also been studied in [13,14], who provide a general convergence result when $(\alpha, \beta, b)$ satisfies certain conditions. However, when $p \geq 2$, the specific choice of $(\alpha, \beta, b)$ in Eq. (4) does not have an analytic form and it thus seems difficult to verify whether $(\alpha, \beta, b)$ in our control system satisfies that condition (see [13, Theorem 2.1])). This topic is beyond the scope of this paper and we leave its investigation to future work.

*Our contribution* Throughout the paper, unless otherwise indicated, we assume that

$\Phi : \mathbb{R}^d \to \mathbb{R}$ *is convex and twice continuously differentiable and the set of global minimizers of* $\Phi$ 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) [6,28,31]. 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:

1. We study the closed-loop control system of Eqs. (3) and (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 [102], showing explicitly that our system extends the high-resolution framework from first-order optimization to high-order optimization.
2. 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 2). 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.
3. 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 [85]. Our iteration complexity analysis is largely motivated by the aforementioned continuous-time analysis, simplifying the analysis in [85] for the case of $p = 2$ and generalizing it to $p > 2$ in a systematic manner (see Theorems 4 and 5 for the details).
4. 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 $\Phi$ which has Lipschitz $p$-th order derivatives. The resulting algorithms include not only existing algorithms studied in [47,61,71] 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 [47,61,71].

*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) [91,95], with a particular focus on the interpretation of Nesterov acceleration as a temporal discretization of a continuous-time dynamical system [3,13,14,17,18,20, 20,21,23,29,30,56,74,83,86,101,102,106,109]. A line of new first-order algorithms have been obtained from the continuous-time dynamics by various advanced numerical integration strategies [40,78,100,103,111,113]. In particular, [100] showed that a basic gradient flow system and multi-step integration scheme yields a class of accelerated first-order optimization algorithms. [113] applied Runge–Kutta integration to an inertial gradient system without Hessian-driven damping [110] 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. [78] and [60] both considered conformal Hamiltonian systems and showed that the resulting discrete-time algorithm achieves fast convergence under certain smoothness conditions. Very recently, [103] have rigorously justified the use of symplectic Euler integrators compared to explicit and implicit Euler integration, which was further studied by [59,87]. 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) [88] and Monteiro and Svaiter's accelerated Newton proximal extragradient (A-NPE) [85]. The ACRN algorithm was extended to a $p$-th order tensor algorithm with the improved convergence rate of $O(k^{-(p+1)})$ [35] and an adaptive $p$-th order tensor algorithm with essentially the same rate [70]. This extension was also revisited by [92] 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 [47,61,71] and was shown to achieve a convergence rate of $O(k^{-(3p+1)/2})$, matching the lower bound [11]. Subsequently, a high-order coordinate descent algorithm was studied in [9], and very recently, the high-order A-NPE framework has been specialized to the strongly convex setting [8], 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 [57,63–66] and an optimal algorithm is known [105]. Other settings include structured convex non-smooth minimization [48], convex-concave minimax optimization and monotone variational inequalities [49,97], and structured smooth convex minimization [72,73,93,94]. In the nonconvex setting, high-order algorithms have been also proposed and analyzed [42,43,50,51,82].

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 [105], and the connection to dynamical systems in that work is unclear. In particular, some aspects of the UAF algorithm (see [105, 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 [78,111,113] which were derived from continuous-time perspective with theoretical guarantee under certain condition. In particular, [111] 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 [47,61,71]. 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.

*Organization* The remainder of the paper is organized as follows. In Sect. 2, we study the closed-loop control system in Eqs. (3) and (4) and prove the existence and uniqueness of a local solution using the Banach fixed-point theorem. In Sect. 3, 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. In Sect. 4, we provide 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. Our iteration complexity analysis is largely motivated by the continuous-time analysis of our system, demonstrating that these algorithms achieve fast gradient minimization. In Sect. 5, we conclude our work with a brief discussion on future research directions.

*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 $\ell_2$ Euclidean norm and let $\mathbb{B}_\delta(x) = \{x' \in \mathbb{R}^d \mid \|x' - x\| \leq \delta\}$ denote its $\delta$-neighborhood. For a tensor $X \in \mathbb{R}^{d_1 \times \cdots \times d_p}$, we define

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

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

Fix $p \geq 1$, we define $\mathcal{F}^p_\ell(\mathbb{R}^d)$ as the class of convex functions on $\mathbb{R}^d$ with $\ell$-Lipschitz $p$-th order derivatives; that is, $f \in \mathcal{F}^p_\ell(\mathbb{R}^d)$ 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

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

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

## 2 The closed-loop control system

In this section, we study the closed-loop control system in Eqs. (3) and (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.

### 2.1 First-order system in time and space

We rewrite the closed-loop control system in Eqs. (3) and (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 $(\alpha, \beta, b)$ explicitly depend on the variables $(x, \lambda, a)$, the parameters $c > 0$, $\theta \in (0, 1)$ and the order $p \in \{1, 2, \ldots\}$:

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

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

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

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

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

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 Eqs. (3) and (4) can be written as a first-order system in time and space as follows:

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

We also provide another first-order system in time and space with different variable $(x, v, \lambda, \gamma)$. 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:

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

***Remark 1*** The first-order systems in Eqs. (7) and (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 2*** The first-order systems in Eqs. (7) and (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 [6,28,31]. In this setting, either the open-loop case or the closed-loop case without inertia has been studied in the literature [1,12,16,17,27,33,34,45,79], but there is significantly less work on the case of a closed-loop control system with inertia. For recent progress in this direction, see [14].

### 2.2 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*

1. $\varphi(\cdot, x)$ *is linear, strictly increasing and* $\varphi(0, x) = 0$.
2. $\varphi(\lambda, x) \to +\infty$ *as* $\lambda \to +\infty$.

***Proof*** By the definition of $\varphi$, 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 $\lambda \to +\infty$. □

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 $\Omega$ and $\Lambda_\theta(\cdot)$ which are crucial to the proof of existence and uniqueness presented in the next subsection.

**Proposition 2** *The set $\Omega$ is open.*

***Proof*** Given $x \in \Omega$, it suffices to show that $\mathbb{B}_\delta(x) \subseteq \Omega$ for some $\delta > 0$. Since $\Phi$ 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. □

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

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

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

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

### 2.3 Existence and uniqueness of a local solution

We prove the existence and uniqueness of a local solution of the closed-loop control system in Eqs. (3) and (4) by appeal to the Banach fixed-point theorem. Using the results in Sect. 2.1 (see Eq. (6)), our system can be equivalently written as follows:

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

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

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

A common method for proving the existence and uniqueness of a local solution is via appeal to the Cauchy-Lipschitz theorem [54, 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 [62], and we avail ourselves directly of the latter theorem. In particular, we construct Picard iterates $\psi_k$ whose limit is a fixed point of a contraction $T$. We have the following theorem.

**Theorem 1** *There exists $t_0 > 0$ such that the autonomous system in Eqs.* (11) *and* (12) *has a unique solution $x : [0, t_0] \mapsto \mathbb{R}^d$.*

***Proof*** 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 $\Phi$ is twice continuously differentiable. By the definition of $\Lambda_\theta$, 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

$$\|F(t, x(t))\| \leq M, \quad \forall t \in [0, 1]. \tag{13}$$

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

$$\|z(t) - x_0\| = \left\| \int_0^t F(s, x(s))ds \right\| \overset{\text{Eq. (13)}}{\leq} Mt \leq Mt_0 \leq Mt_1 \leq \delta.$$

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 $\{\psi_k\}_{k\ge 0}$ with $\psi_0(t) = x_0$ for $\forall t \in [0, t_0]$ and $\psi_{k+1} = T\psi_k$ for all $k \ge 0$. By the Banach fixed-point theorem [62], 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 [107] 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 $\max\{\Lambda_\theta(z), \|\nabla\Phi(z)\|\} \le M_1$ for $\forall z \in \mathbb{B}_\delta(x_0)$. By Proposition 3, $\Lambda_\theta$ and $\sqrt{\Lambda_\theta}$ are continuous and locally Lipschitz over $\Omega$. 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

$$\max\{|\Lambda_\theta(x') - \Lambda_\theta(x'')|, |\sqrt{\Lambda_\theta}(x') - \sqrt{\Lambda_\theta}(x'')|\} \le L_1\|x' - x''\|. \tag{14}$$

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

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

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

The key inequality for the subsequent analysis is as follows:

$$\|a_1b_1 - a_2b_2\| \le \|a_1\|\|b_1 - b_2\| + \|b_2\|\|a_1 - a_2\|. \tag{15}$$

First, by combining Eq. (15) with $\max\{\Lambda_\theta(x(t)), \|\nabla\Phi(x(t))\|\} \le M_1$, $\|\nabla\Phi(x') - \nabla\Phi(x'')\| \le L_2\|x' - x''\|$ and Eq. (14), we obtain:

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

Second, we combine Eq. (15) with $\sqrt{\Lambda_\theta(x(t))} \leq \sqrt{M_1}$, Eq. (14) and $0 < s \leq t_0 < t_1 < 1$ to obtain:

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

We also obtain by combining Eq. (15) 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. (14) and $0 < w \leq s \leq t_0 < t_1 < 1$ that

$$\begin{aligned}&\left\| \int_0^s \left(\sqrt{\Lambda_\theta(x'(w))}\left(\int_0^w \sqrt{\Lambda_\theta(x'(v))}\, dv + c\right)\right) \nabla\Phi(x'(w))dw \right.\\ &\quad \left. - \int_0^s \left(\sqrt{\Lambda_\theta(x''(w))}\left(\int_0^w \sqrt{\Lambda_\theta(x''(v))}\, dv + c\right)\right) \nabla\Phi(x''(w))dw \right\| \\ &\leq (M_1L_2 + c\sqrt{M_1}L_2 + 2(M_1)^{3/2}L_1 + cM_1L_1)d(x', x'').\end{aligned}$$

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

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

Putting these pieces together yields that

$$\mathbf{II} \leq \left(\frac{2(M_1)^{5/2}L_1}{c^2} + \frac{(M_1)^{3/2}L_2 + 5(M_1)^2L_1}{c} + M_1L_2 + 2(M_1)^{3/2}L_1\right) t_0 d(x', x'').$$

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 $c$, $x_0$, $\delta$, $\Phi(\cdot)$ and $\Lambda_\theta(\cdot)$) and is defined as follows:

$$\bar{M} = \frac{2((M_1)^2 + 2M_2 + 2\|v_0\|)\sqrt{M_1}L_1}{c^2} + \frac{2\sqrt{M_1} + (2M_2 + 2\|v_0\| + 5(M_1)^2)L_1 + (M_1)^{3/2}L_2}{c}$$
$$+2M_1L_2 + (M_1 + 2(M_1)^{3/2})L_1.$$

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. □

## 2.4 Discussion

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

*Hessian-driven damping* The formal introduction of Hessian-driven damping in optimization dates to [6], with many subsequent developments; see, e.g., [31]. 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 $\Phi$ and the convergence rate of $o(1/t^2)$ in terms of objective function gap.

Recall also that [102] 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 Eqs. (3) and (4) becomes

$$\ddot{x}(t) + \frac{3}{t}\dot{x}(t) + \theta\nabla^2\Phi(x(t))\dot{x}(t) + \left(\theta + \frac{\theta}{t}\right)\nabla\Phi(x(t)) = 0. \quad (16)$$

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 [13,14]; indeed, in this case $(\alpha, \beta, b)$ in [13, Theorem 2.1] has an analytic form. However, the choice of $(\alpha, \beta, b)$ in our general setting in Eq. (4), for $p \geq 2$, does not have an analytic form and it is difficult to verify whether $(\alpha, \beta, b)$ in this case satisfies their condition.

*Newton and Levenberg–Marquardt regularized systems* The precursor of this perspective was developed by [7] 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, [32] and [34] studied Newton dissipative and Levenberg–Marquardt regularized systems:

$$\begin{aligned} \textbf{(Newton)} \quad & \ddot{x}(t) + \nabla^2 \Phi(x(t))\dot{x}(t) + \nabla \Phi(x(t)) = 0. \\ \textbf{(Levenberg–Marquardt)} \quad & \lambda(t)\dot{x}(t) + \nabla^2 \Phi(x(t))\dot{x}(t) + \nabla \Phi(x(t)) = 0. \end{aligned}$$

These systems have been shown to be well defined and stable with robust asymptotic behavior [1,33,34], further motivating the study of the following inertial gradient system with constant damping and Hessian-driven damping [6]:

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

This system attains strong asymptotic stabilization and fast convergence properties [6,28] and can be extended to solve monotone inclusions with theoretical guarantee [1,12,26,27,33,34,79]. 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 [85].

Recently, [12] 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 [12,33] 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 [12,33] 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 [14] 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 [105,110].

By appealing to a variational formulation, [110] derived the following inertial gradient system with asymptotic vanishing damping:

$$\ddot{x}(t) + \frac{p+2}{t}\dot{x}(t) + C(p+1)^2 t^{p-1}\nabla\Phi(x(t)) = 0. \tag{17}$$

Compared to our closed-loop control system, in Eqs. (3) and (4), the system in Eq. (17) 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, [105] has 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:

$$\ddot{x}(t) + \left(\frac{2\dot{a}(t)}{a(t)} - \frac{\ddot{a}(t)}{\dot{a}(t)}\right)\dot{x}(t) + \left(\frac{(\dot{a}(t))^2}{a(t)}\right)\nabla\Phi(x(t)) = 0. \tag{18}$$

Compared to our closed-loop control system, the system in (18) is open-loop and lacks Hessian-driven damping. Moreover, $a(t)$ needs to be determined by hand and [105] do not establish existence or uniqueness of solutions.

## 3 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.

### 3.1 Existence and uniqueness of a global solution

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

**Theorem 2** *Suppose that $\lambda$ is absolutely continuous on any finite bounded interval. Then the closed-loop control system in Eqs.* (3) *and* (4) *has a unique global solution,* $(x, \lambda, a) : [0, +\infty) \mapsto \mathbb{R}^d \times (0, +\infty) \times (0, +\infty)$.

***Remark 3*** 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 $\lambda$ does not dramatically change. In open-loop Levenberg–Marquardt systems, [34] impose the same condition on the regularization parameters. In closed-loop control systems, however, $\lambda$ is not a given datum but an emergent component of the dynamics. Thus, it is preferable to prove that $\lambda$ satisfies this condition rather than assuming it, as done in [33, Theorem 5.2] and [12, 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 Sect. 3.3 for further discussion.

Recall that the system in Eqs. (3) and (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:

$$\mathcal{E}(t) = a(t)(\Phi(x(t)) - \Phi(x^\star)) + \frac{1}{2}\|v(t) - x^\star\|^2, \tag{19}$$

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

***Remark 4*** Note that the Lyapunov function (19) 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 [1,12,33,34] and the inertial gradient system with asymptotic vanishing damping [31,102,106,112]. In particular, [112] 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 $x$, $\lambda$ and $\nabla\Phi$; see Eq. (7).

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*

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

***Proof*** 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:

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

By the convexity of $\Phi$, we have $\Phi(x(t)) - \Phi(x^\star) - \langle\nabla\Phi(x(t)), x(t) - x^\star\rangle \leq 0$. Since $\dot{a}(t) \geq 0$, we have $\mathbf{I} \leq 0$. 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 $\mathbf{II} \leq -a(t)\theta^{\frac{1}{p}}\|\nabla\Phi(x(t))\|^{\frac{p+1}{p}}$. Putting all these pieces together yields the desired inequality. □

**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*** 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

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

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

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

Note that $\|v(t)-x^\star\| \leq \sqrt{2\mathcal{E}(0)}$ 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:

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

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

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

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

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

Putting these pieces together yields that $\|x(t)-x^\star\| \leq \|x_0-x^\star\|+3\sqrt{\mathcal{E}(0)}$. 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. □

***Proof of Theorem 2:*** 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 Eqs. (3) and (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 Theorem 1).

It remains to show that the maximal solution is a global solution; that is, $T_{\max} = +\infty$, if $\lambda$ is absolutely continuous on any finite bounded interval. Indeed, the property of $\lambda$ 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 Eqs. (3) and (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 Eqs. (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 $\lambda$ 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 Theorem 1 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. □

### 3.2 Rate of convergence

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

**Theorem 3** *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 Eqs.* (3) *and* (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 5*** 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 $\Phi \in \mathcal{F}_\ell^1(\mathbb{R}^d)$ 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 [65,89,102] and its analysis involves different techniques.

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** *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*** 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 $p \geq 2$, 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

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

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

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

By the Hölder inequality, we have

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

Combining these results with the definition of $a$ yields:

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

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

$$\int_0^t \left(\sqrt{a(s)} - \frac{c}{2}\right)^{\frac{2p-2}{3p+1}} ds \leq 2\theta^{-\frac{2}{3p+1}} (\mathcal{E}(0))^{\frac{p-1}{3p+1}} \left(\sqrt{a(t)} - \frac{c}{2}\right)^{\frac{2p+2}{3p+1}}. \quad (20)$$

The remaining steps in the proof are inspired by the Bihari–LaSalle inequality [41,76]. 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. (20) 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. (20) yields that

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

This completes the proof. □

***Proof of Theorem 3*** Since the first-order system in Eq. (7) is equivalent to the closed-loop control system in Eqs. (3) and (4), $(x, \lambda, a) : [0, +\infty) \rightarrow \mathbb{R}^d \times (0, +\infty) \times (0, +\infty)$ 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

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

By Lemma 4, we obtain

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

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

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

This completes the proof. □

### 3.3 Discussion

It is useful to compare our approach to approaches based on time scaling [13,19,21,22] and quasi-gradient methods [14,39].

*Regularity condition* Why is proving the existence and uniqueness of a global solution of the closed-loop control system in Eqs. (3) and (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 $\lambda$ and $\nabla \Phi(x)$; and (iii) the evolution dynamics depends on $\lambda$ 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, $\|\nabla\Phi(x(t))\| \geq \|\nabla\Phi(x(0))\|e^{-t}$ or $\lambda(t) \leq \lambda(0)e^{t}$ can readily be derived for the Levenberg–Marquardt regularized system in [34, Corollary 3.3] and even the closed-loop control systems without inertia in [33, Theorem 5.2] and [12, Theorem 2.4]. Thus, we can not exclude the case of $\lambda(t) \rightarrow +\infty$ 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 [31] 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 [31] studied the following open-loop system with the parameters $(\alpha, \beta)$:

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

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 [23] for the reference). Unfortunately, when $c = 0$ and $p = 1$, the closed-loop control system in Eqs. (3) and (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 [31], leaving doubt as to whether weak convergence holds true for the closed-loop control system in Eqs. (3) and (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 [13,19,21,22]. 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 [21,22], 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 $\alpha$ and $b(\cdot)$. Thus, a clear improvement can be achieved by taking $b(t) \rightarrow +\infty$. This demonstrates the power and potential of time scaling, as further evidenced by recent work on systems with Hessian damping [13] and other systems which are associated with the augmented Lagrangian formulation of the affine constrained convex minimization problem [19].

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 $\lambda$ via $a$ and $\dot{a}$. The time scaling methodology accordingly does not capture the continuous-time interpretation of optimal acceleration in high-order optimization [47,61,71,85]. In contrast, our algebraic equation provides a rigorous justification for the large-step condition in the algorithm of [47,61,71,85] 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 [39] and recently applied by [14] 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 $\nabla E$ remains acute and bounded away from $\frac{\pi}{2}$ (see [36,37,52,53,69] for further geometrical interpretation).

Based on seminal work by [39, Theorem 3.2] and [14, Theorem 7.2], convergence properties for the bounded trajectories of quasi-gradient systems have been established if the function $E$ is KL [44,75]. In [14], the authors considered two closed-loop velocity control systems with a damping potential $\phi$:

$$\ddot{x}(t) + \nabla\phi(\dot{x}(t)) + \nabla\Phi(x(t)) = 0. \tag{21}$$

$$\ddot{x}(t) + \nabla\phi(\dot{x}(t)) + \beta\nabla^2\Phi(x(t))\dot{x}(t) + \nabla\Phi(x(t)) = 0. \tag{22}$$

They proposed to use the Hamiltonian formulation of these systems and accordingly defined a function $E_\lambda$ 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 $\phi$ satisfies some certain growth conditions (see [14, Theorem 7.3 and 9.2]), the systems in Eqs. (21) and (22) both have a quasi-gradient structure for $E_\eta$ for sufficiently small $\eta > 0$. This provides an elegant framework for analyzing the convergence properties of the systems in the form of Eqs. (21) and (22) with specific damping potentials.

Why is analyzing our system hard using the quasi-gradient approach? Our system differs from the systems in Eqs. (21) and (22) in two aspects: (i) the closed-loop control law is designed for the gradient of $\Phi$ rather than the velocity $\dot{x}$; (ii) the damping coefficients are time dependent, depending on $\lambda$ via $a$ and $\dot{a}$, and do not have an analytic form when $p \geq 2$. Considering the first-order systems in Eqs. (7) and (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.

## 4 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 Eqs. (7) and (8). Our approach demonstrates the importance of the large-step condition [85] 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 [92], we derive two specific class of $p$-th order tensor algorithms, one of which recovers existing optimal $p$-th order tensor algorithms [47,61,71] and the other of which leads to a new optimal $p$-th order tensor algorithm.

### 4.1 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,

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

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

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

**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
        $w_{k+1} \in \partial_{\epsilon_{k+1}} \Phi(x_{k+1}),$
        $\|\lambda_{k+1} w_{k+1} + x_{k+1} - \tilde{v}_k\|^2 + 2\lambda_{k+1}\epsilon_{k+1} \leq \sigma^2 \|x_{k+1} - \tilde{v}_k\|^2,$
        $\lambda_{k+1}\|x_{k+1} - \tilde{v}_k\|^{p-1} \geq \theta.$
        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.
```

$$\left\{ \begin{array}{l} v_{k+1} - v_k + a_{k+1} \nabla \Phi(x_{k+1}) = 0 \\ A_{k+1}(x_{k+1} - x_k) + a_{k+1}(x_k - v_k) + \lambda_{k+1} A_{k+1} \nabla \Phi(x_{k+1}) = 0 \\ (a_{k+1})^2 = \lambda_{k+1}(A_k + a_{k+1}), \; a_{k+1} = A_{k+1} - A_k, \; a_0 = 0 \\ (\lambda_{k+1})^p \|\nabla \Phi(x_{k+1})\|^{p-1} = \theta. \end{array} \right. \tag{23}$$

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. (23) 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. (23). 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 [84,104]. In particular, we define the $\varepsilon$-subdifferential of a function $f$ by

$$\partial_\epsilon f(x) := \{w \in \mathbb{R}^d \mid f(y) \geq f(x) + \langle y - x, w \rangle - \epsilon, \; \forall y \in \mathbb{R}^d\}, \tag{24}$$

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 6*** We present our first conceptual algorithmic framework formally in Algorithm 1. This scheme includes the large-step A-HPE framework [85] 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 [85].

*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 $c = 2$:

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

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

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

$$\left\{\begin{array}{l} v_{k+1} - v_k + \frac{\alpha_{k+1}}{\gamma_{k+1}}\nabla\Phi(x_{k+1}) = 0 \\ x_{k+1} - x_k + \alpha_{k+1}(x_k - v_k) + \lambda_{k+1}\nabla\Phi(x_{k+1}) = 0 \\ (\alpha_{k+1})^2 = \lambda_{k+1}\gamma_{k+1}, \ \gamma_{k+1} = (1 - \alpha_{k+1})\gamma_k, \ \gamma_0 = 1 \\ (\lambda_{k+1})^p\|\nabla\Phi(x_{k+1})\|^{p-1} = \theta. \end{array}\right. \quad (25)$$

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. (23) 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. (25).

***Remark 7*** 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 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) [67], which combines Nes-

**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
    $w_{k+1} \in \partial_{\epsilon_{k+1}} \Phi(x_{k+1})$,
    $\|\lambda_{k+1} w_{k+1} + x_{k+1} - \tilde{v}_k\|^2 + 2\lambda_{k+1}\epsilon_{k+1} \leq \sigma^2 \|x_{k+1} - \tilde{v}_k\|^2$,
    $\lambda_{k+1} \|x_{k+1} - \tilde{v}_k\|^{p-1} \geq \theta$.
  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.
```

terov acceleration [95] and Martinet's PPA [80,81]. 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 $\Phi$ via the large-step condition.

Such difference is emphasized by recent studies on the continuous-time perspective of Güler's APPA [21,22]. More specifically, [21] proved that Güler's APPA can be interpreted as the implicit time discretization of an open-loop inertial gradient system (see [21, 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 $\beta_k$ in their notation correspond to $\alpha_k$ and $\lambda_k$ 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

$$\ddot{x}(t) + \left(\frac{2\dot{a}(t)}{a(t)} - \frac{\ddot{a}(t)}{\dot{a}(t)}\right)\dot{x}(t) + \left(\frac{(\dot{a}(t))^2}{a(t)}\right)\nabla\Phi(x(t)) = 0, \tag{26}$$

where $a$ explicitly depends on the variable $\lambda$ as follows,

$$a(t) = \frac{1}{4}\left(\int_0^t \sqrt{\lambda(s)}ds + 2\right)^2.$$

Compared to our closed-loop control system, the one in Eq. (26) 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 [13].

## 4.2 Complexity analysis

We study the iteration complexity of Algorithms 1 and 2. Our analysis is largely motivated by the aforementioned continuous-time analysis, simplifying the analysis in [85] for the case of $p = 2$ and generalizing it to the case of $p > 2$ in a systematic manner (see Theorems 4 and 5). Throughout this subsection, $x^\star$ denotes the projection of $v_0$ onto the solution set of $\Phi$.

*Algorithm* 1 We start with the presentation of our main results for Algorithm 1, which generalizes [85, Theorem 4.1] to the case of $p > 2$.

**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*

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

Note that the only difference between Algorithm 1 and large-step A-HPE framework in [85] is the order in the algebraic equation. Thus, many of the technical results derived in [85] also hold for Algorithm 1; more specifically, [85, Theorem 3.6, Lemma 3.7 and Proposition 3.9].

We also 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*

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

***Proof*** For $p = 1$, the large-step condition implies that $\lambda_k \geq \theta$ for all $k \geq 0$. By [85, Lemma 3.7], we have $A_k \geq \frac{\theta k^2}{4}$.

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

$$\begin{aligned}\sum_{i=1}^k A_i(\lambda_i)^{-\frac{p+1}{p-1}}\theta^{\frac{2}{p-1}} &\leq \sum_{i=1}^k A_i(\lambda_i)^{-\frac{p+1}{p-1}}(\lambda_i\|x_i - \tilde{v}_{i-1}\|^{p-1})^{\frac{2}{p-1}} \\ &= \sum_{i=1}^k \frac{A_i}{\lambda_i}\|x_i - \tilde{v}_{i-1}\|^2 \overset{[85, Theorem\ 3.6]}{\leq} \frac{\|v_0 - x^\star\|^2}{1-\sigma^2}.\end{aligned}$$

By the Hölder inequality, we have

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

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:

$$\sum_{i=1}^{k}(A_i)^{\frac{p-1}{3p+1}} \leq C\left(\sum_{i=1}^{k}\sqrt{\lambda_i}\right)^{\frac{2p+2}{3p+1}} \overset{[85, Lemma\ 3.7]}{\leq} 2C(A_k)^{\frac{p+1}{3p+1}}. \tag{27}$$

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. (27) implies that

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

This implies that

$$y_k - y_{k-1} \geq \left(\frac{y_k}{2C}\right)^{\frac{p-1}{p+1}} \Longrightarrow \frac{y_k - y_{k-1}}{(y_k)^{\frac{p-1}{p+1}}} \geq \left(\frac{1}{2C}\right)^{\frac{p-1}{p+1}}. \tag{28}$$

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

$$(y_k)^{\frac{2}{p+1}} - (y_{k-1})^{\frac{2}{p+1}} \geq \frac{2}{p+1}\left(\frac{y_k - y_{k-1}}{[y_k]^{\frac{p-1}{p+1}}}\right). \tag{29}$$

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. (28) follows from setting $t = \frac{y_{k-1}}{y_k}$. Combining Eqs. (28) and (29) 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. (27) 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. □

***Remark 8*** The proof of Lemma 5 is much simpler than the existing analysis; e.g., [85, Lemma 4.2] for the case of $p = 2$ and [71, Theorem 3.4] and [47, Lemma 3.3] for the case of $p \geq 2$. Notably, it is not a generalization of the highly technical proof in [85, Lemma 4.2] but can be interpreted as the discrete-time counterpart of the proof of Lemma 4.

***Proof of Theorem 4:*** For every integer $k \geq 1$, by [85, 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 [85, Proposition 3.9] and Lemma 5, we have

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

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

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

This completes the proof. □

*Algorithm* 2 We now present our main results for Algorithm 2. The proof is analogous to that of Theorem 4 and based on another estimate sequence.

**Theorem 5** *For every integer $k \geq 1$, the objective function gap satisfies*

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

*and*

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

Inspired by the continuous-time Lyapunov function in Eq. (19), we construct a discrete-time Lypanunov function for Algorithm 2 as follows:

$$\mathcal{E}_k = \frac{1}{\gamma_k}(\Phi(x_k) - \Phi(x^\star)) + \frac{1}{2}\|v_k - x^\star\|^2. \tag{30}$$

We use this function to prove technical results that pertain to Algorithm 2 and which are the analogs of [85, Theorem 3.6, Lemma 3.7 and Proposition 3.9].

**Lemma 6** *For every integer $k \geq 1$,*

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

*which implies that*

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

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

***Proof*** 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 ($\Gamma_k$ 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, $\phi_k$ is affine. Since $w_{k+1} \in \partial_{\epsilon_{k+1}} \Phi(x_{k+1})$, Eq. (24) implies that $\phi_k(v) \leq \Phi(v) - \Phi(x^\star)$. Furthermore, $\Gamma_k$ is quadratic and $\nabla^2 \Gamma_k = \nabla^2 \Gamma_0$ since $\phi_k$ 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

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

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 $\Gamma_k$ 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 $k \geq 0$,

$$\beta_{k+1} \geq \beta_k + \frac{1-\sigma^2}{2\lambda_{k+1}\gamma_{k+1}} \|x_{k+1} - \tilde{v}_k\|^2. \tag{31}$$

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 $\phi_{k+1}$, we have

$$\phi_{k+1}(\tilde{v}) = (1 - \alpha_{k+1})\phi_{k+1}(x_k) + \alpha_{k+1}\phi_{k+1}(v), \tag{32}$$

$$\tilde{v} - \tilde{v}_k = \alpha_{k+1}(v - v_k). \tag{33}$$

Since $\Gamma_k$ 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 $\Gamma_k$ 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 $\beta_k$, 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

$$\begin{aligned}\Gamma_{k+1}(v) &\geq \beta_k + \frac{\alpha_{k+1}}{\gamma_{k+1}}\phi_{k+1}(v) + \frac{1}{\gamma_k}\phi_{k+1}(x_k) + \frac{1}{2}\|v - v_k\|^2 \\ &\overset{\text{Eq. (32)}}{=} \beta_k + \frac{1}{\gamma_{k+1}}\phi_{k+1}(\tilde{v}) + \frac{1}{2}\|v - v_k\|^2 \\ &= \beta_k + \frac{1}{\gamma_{k+1}}\left(\phi_{k+1}(\tilde{v}) + \frac{\gamma_{k+1}}{2}\|v - v_k\|^2\right) \\ &\overset{\text{Eq. (33)}}{=} \beta_k + \frac{1}{\gamma_{k+1}}\left(\phi_{k+1}(\tilde{v}) + \frac{\gamma_{k+1}}{2(\alpha_{k+1})^2}\|\tilde{v} - \tilde{v}_k\|^2\right) \\ &= \beta_k + \frac{1}{\gamma_{k+1}}\left(\phi_{k+1}(\tilde{v}) + \frac{1}{2\lambda_{k+1}}\|\tilde{v} - \tilde{v}_k\|^2\right).\end{aligned}$$

Using [85, 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 $\beta_k$ yields the desired inequality in Eq. (31). This completes the proof. □

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

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

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

***Proof*** 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. □

**Lemma 8** *For every integer $k \geq 1$ and $\sigma < 1$, there exists $1 \leq i \leq k$ such that*

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

***Proof*** 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 $\tau_k$ yields the desired results. □

As the analog of Lemma 5, we provide a technical lemma on the upper bound for $\gamma_k$. 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*** 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

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

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

$$\sum_{i=1}^{k}(\gamma_i)^{-\frac{p-1}{3p+1}} \leq C\left(\sum_{i=1}^{k}\sqrt{\lambda_i}\right)^{\frac{2p+2}{3p+1}} \overset{\text{Lemma 7}}{\leq} 2C(\gamma_k)^{-\frac{p+1}{3p+1}}. \tag{34}$$

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. (34) 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. □

***Proof of Theorem 5:*** For every integer $k \geq 1$, by Lemmas 6 and 9, we have

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

By Lemmas 8 and 9, we have

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

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

As in the proof of Theorem 4, we conclude that $\inf_{1 \le i \le k} \|w_i\|^2 = O(k^{-3p})$. This completes the proof. □

***Remark 9*** The discrete-time analysis in this subsection is based on a discrete-time Lyapunov function in Eq. (30), which is closely related to the continuous one in Eq. (19), and two simple yet nontrivial technical lemmas (see Lemmas 5 and 9 ), which are both discrete-time versions of Lemma 4. Notably, the proofs of Lemmas 5 and 9 follows the same path for proving Lemma 4 and have demanded the use of the Bihari–LaSalle inequality in discrete time.

### 4.3 Optimal tensor algorithms and gradient norm minimization

By instantiating Algorithms 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 in [47,61,71] 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 [85,102].

*Approximate tensor subroutine* Proximal point algorithms [67,99] (corresponding to implicit time discretization of certain systems) require solving an exact proximal iteration with proximal coefficient $\lambda > 0$ at each iteration:

$$x = \operatorname*{argmin}_{u \in \mathbb{R}^d} \left\{ \Phi(u) + \frac{1}{2\lambda} \|u - v\|^2 \right\}. \tag{35}$$

In general, Eq. (35) can be as hard as minimizing the function $\Phi$ 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 $\Phi$ and a regularization term, motivating a line of $p$-th order tensor algorithms [35,42,43,47,61,70,71,82,92]. More specifically, we define

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

**Algorithm 3** Optimal $p$-th order Tensor Algorithm I [47,61,71]

```
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 $\lambda_{k+1}$ with a $\hat{\sigma}$-inexact solution $x_{k+1} \in \mathbb{R}^d$ of Eq. (36a)
satisfying that
    $\frac{\sigma_l p!}{2\ell} \leq \lambda_{k+1} \left\|x_{k+1} - \tilde{v}_k\right\|^{p-1} \leq \frac{\sigma_u p!}{2\ell}$,
or an exact solution $x_{k+1} \in \mathbb{R}^d$ of Eq. (36b) satisfying that
    $\frac{(p-1)!}{2\ell} \leq \lambda_{k+1} \left\|x_{k+1} - \tilde{v}_k\right\|^{p-1} \leq \frac{p!}{\ell(p+1)}$,
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.
```

The algorithms of this subsection are based on either an inexact solution of Eq. (36a), used in [71], or an exact solution of Eq. (36b), used in [47,61]:

$$\min_{u \in \mathbb{R}^d} \ \Phi_v(u) + \frac{1}{2\lambda}\|u - v\|^2, \tag{36a}$$

$$\min_{u \in \mathbb{R}^d} \ \Phi_v(u). \tag{36b}$$

In particular, the solution $x_v$ of Eq. (36a) is unique and satisfies $\lambda \nabla \Phi_v(x_v) + x_v - v = 0$. Thus, we denote a $\hat{\sigma}$*-inexact solution* of Eq. (36a) 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. (36b) 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 $\theta$.

**Proposition 4** *Algorithm 3 is Algorithm 1 with* $\theta = \frac{\sigma_l p!}{2\ell}$ *or* $\theta = \frac{(p-1)!}{2\ell}$.

***Proof*** 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 [71, Proposition 3.2] with a $\hat{\sigma}$-inexact solution $x_{k+1} \in \mathbb{R}^d$ of Eq. (36a) 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

$$\lambda_{k+1}\|x_{k+1} - \tilde{v}_k\|^{p-1} \leq \frac{\sigma_u p!}{2\ell} \Longrightarrow \hat{\sigma} + \frac{2\ell\lambda_{k+1}}{p!}\|x_{k+1} - \tilde{v}_k\|^{p-1} \leq \hat{\sigma} + \sigma_u = \sigma,$$

$$\lambda_{k+1}\|x_{k+1} - \tilde{v}_k\|^{p-1} \geq \frac{\sigma_l p!}{2\ell} \Longrightarrow \lambda_{k+1}\|x_{k+1} - \tilde{v}_k\|^{p-1} \geq \theta.$$

**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 $\lambda_{k+1}$ with a $\hat{\sigma}$-inexact solution $x_{k+1} \in \mathbb{R}^d$ of Eq. (36a) 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. (36b) 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.
```

Using the same argument with [47, Lemma 3.1] instead of [71, Proposition 3.2] and an exact solution $x_{k+1} \in \mathbb{R}^d$ of Eq. (36b), we obtain the same result with $\theta = \frac{(p-1)!}{2\ell}$. Putting these pieces together yields the desired conclusion. □

In view of Proposition 4, the iteration complexity derived for Algorithm 1 hold for Algorithm 3. We summarize the results in the following theorem.

**Theorem 6** *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 10*** Theorem 6 has been derived in [85, 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 [102]. For $p \geq 3$ in general, the first inequality on the objective function gap has been derived independently in [61, Theorem 1], [71, Theorem 3.5] and [47, 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 $\theta$. The proof is omitted since it is the same as the aforementioned analysis for Algorithm 3.

**Proposition 5** *Algorithm 4 is Algorithm 2 with $\theta = \frac{\sigma_l p!}{2\ell}$ or $\theta = \frac{(p-1)!}{2\ell}$.*

**Theorem 7** *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 11*** The approximate tensor subroutine in Algorithms 3 and 4 can be efficiently implemented usinga novel bisection search scheme. We refer the interested readers to [47,71] for the details.

## 5 Conclusions

We have presented a closed-loop control system for modeling optimal tensor algorithms for smooth convex optimization and provided continuous-time and discrete-time Lyapunov functions for analyzing the convergence properties of this system and its discretization. Our framework provides a systematic way to derive discrete-time $p$-th order optimal tensor algorithms, for $p \geq 2$, and simplify existing analyses via the use of a Lyapunov function. A key ingredient in our framework is the algebraic equation, which is not present in the setting of $p = 1$, but is essential for deriving optimal acceleration methods for $p \geq 2$. Our framework allows us to infer that a certain class of $p$-th order tensor algorithms minimize the squared norm of the gradient at a fast rate of $O(k^{-3p})$ for smooth convex functions.

It is worth noting that one could also consider closed-loop feedback control of the velocity. This is called nonlinear damping in the PDE literature; see [14] for recent progress in this direction. There are also several other avenues for future research. In particular, it is of interest to bring our perspective into register with the Lagrangian and Hamiltonian frameworks that have proved productive in recent work [55,59,87,110], as well as the control-theoretic viewpoint of [68,77]. We would hope for this study to provide additional insight into the geometric or dynamical role played by the algebraic equation for modeling the continuous-time dynamics. Moreover, we wish to study possible extensions of our framework to nonsmooth optimization by using differential inclusions [109] and monotone inclusions. The idea is to consider the setting in which $0 \in T(x)$ where $T$ is a maximally monotone operator in a Hilbert space [1,5,12,16,17,26,27,33,34,45,79]. Finally, given that we know that direct discretization of our closed-loop control system cannot recover Nesterov's optimal high-order tensor algorithms [91, Section 4.3], it is of interest to investigate the continuous-time limit of Nesterov's algorithms and see whether the algebraic equation plays a role in their analysis.

**Acknowledgements** The authors would like to thank co-editor Prof. Adrian S. Lewis, the associate editor and two anonymous reviewers for constructive comments that improve the presentation of this paper. This work was supported in part by the Mathematical Data Science program of the Office of Naval Research under grant number N00014-18-1-2764.

## References

1. Abbas, B., Attouch, H., Svaiter, B.F.: Newton-like dynamics and forward–backward methods for structured monotone inclusions in Hilbert spaces. J. Optim. Theory Appl. **161**(2), 331–360 (2014)
2. Adly, S., Attouch, H.: Finite convergence of proximal-gradient inertial algorithms combining dry friction with hessian-driven damping. SIAM J. Optim. **30**(3), 2134–2162 (2020)
3. Adly, S., Attouch, H.: First-order inertial algorithms involving dry friction damping. Math. Program. 1–41 (2021)
4. Alvarez, F.: On the minimizing property of a second order dissipative system in Hilbert spaces. SIAM J. Control Optim. **38**(4), 1102–1119 (2000)
5. Alvarez, F., Attouch, H.: An inertial proximal method for maximal monotone operators via discretization of a nonlinear oscillator with damping. Set Valued Anal. **9**(1), 3–11 (2001)
6. Alvarez, F., Attouch, H., Bolte, J., Redont, P.: 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)
7. Alvarez, F., Pérez C, J.M.: A dynamical system associated with Newton's method for parametric approximations of convex minimization problems. Appl. Math. Optim. **38**, 193–217 (1998)
8. Alves, M.M.: 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: arXiv:2102.02045 (2021)
9. Amaral, V.S., Andreani, R., Birgin, E.G., Marcondes, D.S., Martínez, J.M.: On complexity and convergence of high-order coordinate descent algorithms. ArXiv Preprint: arXiv:2009.01811 (2020)
10. Antipin, A.S.: Minimization of convex functions on convex sets by means of differential equations. Differ. Equ. **30**(9), 1365–1375 (1994)
11. Arjevani, Y., Shamir, O., Shiff, R.: Oracle complexity of second-order methods for smooth convex optimization. Math. Program. **178**(1), 327–360 (2019)
12. Attouch, H., Alves, M.M., Svaiter, B.F.: A dynamic approach to a proximal-Newton method for monotone inclusions in Hilbert spaces, with complexity o (1/nˆ2). J. Convex Anal. **23**(1), 139–180 (2016)
13. Attouch, H., Balhag, A., Chbani, Z., Riahi, H.: Fast convex optimization via inertial dynamics combining viscous and Hessian-driven damping with time rescaling. Evol. Equ. Control Theory (to appear) (2021)
14. Attouch, H., Bot, R.I., Csetnek, E.R.: Fast optimization via inertial dynamics with closed-loop damping. ArXiv Preprint: arXiv:2008.02261 (2020)
15. Attouch, H., Cabot, A.: Asymptotic stabilization of inertial gradient dynamics with time-dependent viscosity. J. Differ. Equ. **263**(9), 5412–5458 (2017)
16. Attouch, H., Cabot, A.: Convergence of damped inertial dynamics governed by regularized maximally monotone operators. J. Differ. Equ. **264**(12), 7138–7182 (2018)
17. Attouch, H., Cabot, A.: Convergence of a relaxed inertial proximal algorithm for maximally monotone operators. Math. Program. **184**(1), 243–287 (2020)
18. Attouch, H., Chbani, Z., Fadili, J., Riahi, H.: First-order optimization algorithms via inertial systems with Hessian driven damping. Math. Program. 1–43 (2020)
19. Attouch, H., Chbani, Z., Fadili, J., Riahi, H.: Fast convergence of dynamical ADMM via time scaling of damped inertial dynamics. ArXiv Preprint: arXiv:2103.12675 (2021)
20. Attouch, H., Chbani, Z., Peypouquet, J., Redont, P.: Fast convergence of inertial dynamics and algorithms with asymptotic vanishing viscosity. Math. Program. **168**(1–2), 123–175 (2018)

21. Attouch, H., Chbani, Z., Riahi, H.: Fast convex optimization via time scaling of damped inertial gradient dynamics. Pure Appl. Funct. Anal. (to appear) (2019)
22. Attouch, H., Chbani, Z., Riahi, H.: Fast proximal methods via time scaling of damped inertial dynamics. SIAM J. Optim. **29**(3), 2227–2256 (2019)
23. Attouch, H., Chbani, Z., Riahi, H.: Rate of convergence of the Nesterov accelerated gradient method in the subcritical case $alphale3$. ESAIM Control Optim. Calc. Var. **25**, 2 (2019)
24. Attouch, H., Cominetti, R.: A dynamical approach to convex minimization coupling approximation with the steepest descent method. J. Differ. Equ. **128**(2), 519–540 (1996)
25. Attouch, H., Goudou, X., Redont, P.: 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. Commun. Contemp. Math. **2**(01), 1–34 (2000)
26. Attouch, H., László, S.C.: Continuous Newton-like inertial dynamics for monotone inclusions. Set Valued Var. Anal. 1–27 (2020)
27. Attouch, H., László, S.C.: Newton-like inertial dynamics and proximal algorithms governed by maximally monotone operators. SIAM J. Optim. **30**(4), 3252–3283 (2020)
28. Attouch, H., Maingé, P.E., Redont, P.: A second-order differential system with Hessian-driven damping: application to non-elastic shock laws. Differ. Equ. Appl. **4**(1), 27–65 (2012)
29. Attouch, H., Peypouquet, J.: The rate of convergence of Nesterov's accelerated forward-backward method is actually faster than 1/kˆ2. SIAM J. Optim. **26**(3), 1824–1834 (2016)
30. Attouch, H., Peypouquet, J.: Convergence rate of proximal inertial algorithms associated with Moreau envelopes of convex functions. In: Splitting Algorithms, Modern Operator Theory, and Applications, pp. 1–44. Springer (2019)
31. Attouch, H., Peypouquet, J., Redont, P.: Fast convex optimization via inertial dynamics with Hessian driven damping. J. Differ. Equ. **261**(10), 5734–5783 (2016)
32. Attouch, H., Redont, P.: The second-order in time continuous Newton method. In: Approximation, Optimization and Mathematical Economics, pp. 25–36. Springer (2001)
33. Attouch, H., Redont, P., Svaiter, B.F.: Global convergence of a closed-loop regularized Newton method for solving monotone inclusions in Hilbert spaces. J. Optim. Theory Appl. **157**(3), 624–650 (2013)
34. Attouch, H., Svaiter, B.F.: A continuous dynamical Newton-like approach to solving monotone inclusions. SIAM J. Control Optim. **49**(2), 574–598 (2011)
35. Baes, M.: Estimate Sequence Methods: Extensions and Approximations. Institute for Operations Research, ETH, Zürich (2009)
36. Bárta, T., Chill, R., Fašangová, E.: Every ordinary differential equation with a strict Lyapunov function is a gradient system. Monatshefte für Mathematik **166**(1), 57–72 (2012)
37. Bárta, T., Fašangová, E.: Convergence to equilibrium for solutions of an abstract wave equation with general damping function. J. Differ. Equ. **260**(3), 2259–2274 (2016)
38. Beck, A., Teboulle, M.: A fast iterative shrinkage-thresholding algorithm for linear inverse problems. SIAM J. Imaging Sci. **2**(1), 183–202 (2009)
39. Bégout, P., Bolte, J., Jendoubi, M.A.: On damped second-order gradient systems. J. Differ. Equ. **259**(7), 3115–3143 (2015)
40. Betancourt, M., Jordan, M.I., Wilson, A.C.: On symplectic optimization. ArXiv Preprint: arXiv:1802.03653 (2018)
41. Bihari, I.: A generalization of a lemma of Bellman and its application to uniqueness problems of differential equations. Acta Mathematica Hungarica **7**(1), 81–94 (1956)
42. Birgin, E.G., Gardenghi, J.L., Martinez, J.M., Santos, S.A., Toint, P.L.: Evaluation complexity for nonlinear constrained optimization using unscaled KKT conditions and high-order models. SIAM J. Optim. **26**(2), 951–967 (2016)
43. Birgin, E.G., Gardenghi, J.L., Martínez, J.M., Santos, S.A., Toint, P.L.: Worst-case evaluation complexity for unconstrained nonlinear optimization using high-order regularized models. Math. Program. **163**(1–2), 359–368 (2017)
44. Bolte, J., Daniilidis, A., Ley, O., Mazet, L.: Characterizations of łojasiewicz inequalities: subgradient flows, talweg, convexity. Trans. Am. Math. Soc. **362**(6), 3319–3363 (2010)
45. Bot, R.I., Csetnek, E.R.: Second order forward–backward dynamical systems for monotone inclusion problems. SIAM J. Control Optim. **54**(3), 1423–1443 (2016)
46. Boţ, R.I., Csetnek, E.R., László, S.C.: Tikhonov regularization of a second order dynamical system with Hessian driven damping. Math. Program. 1–36 (2020)

47. Bubeck, S., Jiang, Q., Lee, Y.T., Li, Y., Sidford, A.: Near-optimal method for highly smooth convex optimization. In: COLT, pp. 492–507. PMLR (2019)
48. Bullins, B.: Highly smooth minimization of nonsmooth problems. In: COLT, pp. 988–1030. PMLR (2020)
49. Bullins, B., Lai, K.A.: Higher-order methods for convex–concave min–max optimization and monotone variational inequalities. ArXiv Preprint: arXiv:2007.04528 (2020)
50. Cartis, C., Gould, N.I., Toint, P.L.: Universal regularization methods: varying the power, the smoothness and the accuracy. SIAM J. Optim. **29**(1), 595–615 (2019)
51. Cartis, C., Gould, N.I.M., Toint, P.L.: Second-order optimality and beyond: characterization and evaluation complexity in convexly constrained nonlinear optimization. Found. Comput. Math. **18**(5), 1073–1107 (2018)
52. Chergui, L.: Convergence of global and bounded solutions of a second order gradient like system with nonlinear dissipation and analytic nonlinearity. J. Dyn. Differ. Equ. **3**(20), 643–652 (2008)
53. Chill, R., Fašangová, E.: Gradient systems. In: Lecture Notes of the 13th International Internet Seminar. Matfyzpress, Prague (2010)
54. Coddington, E.A., Levinson, N.: Theory of Ordinary Differential Equations. Tata McGraw-Hill Education, New York (1955)
55. Diakonikolas, J., Jordan, M.I.: Generalized momentum-based methods: a Hamiltonian perspective. SIAM J. Optim. (to appear) (2020)
56. Diakonikolas, J., Orecchia, L.: The approximate duality gap technique: a unified theory of first-order methods. SIAM J. Optim. **29**(1), 660–689 (2019)
57. Doikov, N., Nesterov, Y.: Local convergence of tensor methods. Math. Program. 1–22 (2019)
58. Fazlyab, M., Ribeiro, A., Morari, M., Preciado, V.M.: Analysis of optimization algorithms via integral quadratic constraints: nonstrongly convex problems. SIAM J. Optim. **28**(3), 2654–2689 (2018)
59. França, G., Jordan, M.I., Vidal, R.: On dissipative symplectic integration with applications to gradient-based optimization. J. Stat. Mech. Theory Exp. (to appear) (2021)
60. França, G., Sulam, J., Robinson, D.P., Vidal, R.: Conformal symplectic and relativistic optimization. J. Stat. Mech. Theory Exp. **2020**(12), 124008 (2020)
61. Gasnikov, A., Dvurechensky, P., Gorbunov, E., Vorontsova, E., Selikhanovych, D., Uribe, C.A.: Optimal tensor methods in smooth convex and uniformly convex optimization. In: COLT, pp. 1374–1391. PMLR (2019)
62. Granas, A., Dugundji, J.: Fixed Point Theory. Springer, Berlin (2013)
63. Grapiglia, G.N., Nesterov, Y.: Regularized Newton methods for minimizing functions with Hölder continuous Hessians. SIAM J. Optim. **27**(1), 478–506 (2017)
64. Grapiglia, G.N., Nesterov, Y.: Accelerated regularized Newton methods for minimizing composite convex functions. SIAM J. Optim. **29**(1), 77–99 (2019)
65. Grapiglia, G.N., Nesterov, Y.: Tensor methods for finding approximate stationary points of convex functions. Optim. Methods Softw. 1–34 (2020)
66. Grapiglia, G.N., Nesterov, Y.: Tensor methods for minimizing convex functions with Hölder continuous higher-order derivatives. SIAM J. Optim. **30**(4), 2750–2779 (2020)
67. Güler, O.: New proximal point algorithms for convex minimization. SIAM J. Optim. **2**(4), 649–664 (1992)
68. Hu, B., Lessard, L.: Dissipativity theory for Nesterov's accelerated method. In: ICML, pp. 1549–1557. JMLR. org (2017)
69. Huang, S.Z.: Gradient Inequalities: With Applications to Asymptotic Behavior and Stability of Gradient-Like Systems, vol. 126. American Mathematical Soc, Providence (2006)
70. Jiang, B., Lin, T., Zhang, S.: A unified adaptive tensor approximation scheme to accelerate composite convex optimization. SIAM J. Optim. **30**(4), 2897–2926 (2020)
71. Jiang, B., Wang, H., Zhang, S.: An optimal high-order tensor method for convex optimization. In: COLT, pp. 1799–1801. PMLR (2019)
72. Kamzolov, D.: Near-optimal hyperfast second-order method for convex optimization. In: International Conference on Mathematical Optimization Theory and Operations Research, pp. 167–178. Springer (2020)
73. Kamzolov, D., Gasnikov, A.: Near-optimal hyperfast second-order method for convex optimization and its sliding. ArXiv Preprint: arXiv:2002.09050 (2020)
74. Krichene, W., Bayen, A., Bartlett, P.L.: Accelerated mirror descent in continuous and discrete time. In: NeurIPS, pp. 2845–2853 (2015)

75. Kurdyka, K.: On gradients of functions definable in o-minimal structures. In: Annales de l'institut Fourier **48**, 769–783 (1998)
76. LaSalle, J.: Uniqueness theorems and successive approximations. Ann. Math. **50**, 722–730 (1949)
77. Lessard, L., Recht, B., Packard, A.: Analysis and design of optimization algorithms via integral quadratic constraints. SIAM J. Optim. **26**(1), 57–95 (2016)
78. Maddison, C.J., Paulin, D., Teh, Y.W., O'Donoghue, B., Doucet, A.: Hamiltonian descent methods. ArXiv Preprint: arXiv:1809.05042 (2018)
79. Maingé, P.E.: First-order continuous Newton-like systems for monotone inclusions. SIAM J. Control Optim. **51**(2), 1615–1638 (2013)
80. Martinet, B.: Régularisation d'inéquations variationnelles par approximations successives. rev. française informat. Recherche Opérationnelle **4**, 154–158 (1970)
81. Martinet, B.: Détermination approchée d'un point fixe d'une application pseudo-contractante. CR Acad. Sci. Paris **274**(2), 163–165 (1972)
82. Martínez, J.: On high-order model regularization for constrained optimization. SIAM J. Optim. **27**(4), 2447–2458 (2017)
83. May, R.: Asymptotic for a second-order evolution equation with convex potential and vanishing damping term. Turk. J. Math. **41**(3), 681–685 (2017)
84. Monteiro, R.D.C., Svaiter, B.F.: On the complexity of the hybrid proximal extragradient method for the iterates and the ergodic mean. SIAM J. Optim. **20**(6), 2755–2787 (2010)
85. Monteiro, R.D.C., Svaiter, B.F.: An accelerated hybrid proximal extragradient method for convex optimization and its implications to second-order methods. SIAM J. Optim. **23**(2), 1092–1125 (2013)
86. Muehlebach, M., Jordan, M.I.: A dynamical systems perspective on Nesterov acceleration. In: ICML, pp. 4656–4662 (2019)
87. Muehlebach, M., Jordan, M.I.: Optimization with momentum: dynamical, control-theoretic, and symplectic perspectives. J. Mach. Learn. Res. (to appear) (2021)
88. Nesterov, Y.: Accelerating the cubic regularization of Newton's method on convex problems. Math. Program. **112**(1), 159–181 (2008)
89. Nesterov, Y.: How to make the gradients small. Optima **88**, 10–11 (2012)
90. Nesterov, Y.: Gradient methods for minimizing composite functions. Math. Program. **140**(1), 125–161 (2013)
91. Nesterov, Y.: Lectures on Convex Optimization, vol. 137. Springer, Berlin (2018)
92. Nesterov, Y.: Implementable tensor methods in unconstrained convex optimization. Math. Program. 1–27 (2019)
93. Nesterov, Y.: Inexact accelerated high-order proximal-point methods. Université catholique de Louvain, Center for Operations Research and Econometrics (CORE), Technical report (2020)
94. Nesterov, Y.: Superfast second-order methods for unconstrained convex optimization. Université catholique de Louvain, Center for Operations Research and Econometrics (CORE), Technical report (2020)
95. Nesterov, Y.E.: A method for solving the convex programming problem with convergence rate o $(1/extk, hat, 2)$. Dokl. akad. nauk Sssr **269**, 543–547 (1983)
96. O'Donoghue, B., Maddison, C.J.: Hamiltonian descent for composite objectives. In: NeurIPS, pp. 14470–14480 (2019)
97. Ostroukhov, P., Kamalov, R., Dvurechensky, P., Gasnikov, A.: Tensor methods for strongly convex strongly concave saddle point problems and strongly monotone variational inequalities. ArXiv Preprint: arXiv:2012.15595 (2020)
98. Polyak, B.T.: Introduction to Optimization. Optimization Software Inc, New York (1987)
99. Rockafellar, R.T.: Monotone operators and the proximal point algorithm. SIAM J. Control Optim. **14**(5), 877–898 (1976)
100. Scieur, D., Roulet, V., Bach, F., d'Aspremont, A.: Integration methods and optimization algorithms. In: NeurIPS, pp. 1109–1118 (2017)
101. Sebbouh, O., Dossal, C., Rondepierre, A.: Convergence rates of damped inertial dynamics under geometric conditions and perturbations. SIAM J. Optim. **30**(3), 1850–1877 (2020)
102. Shi, B., Du, S.S., Jordan, M.I., Su, W.J.: Understanding the acceleration phenomenon via high-resolution differential equations. ArXiv Preprint: arXiv:1810.08907 (2018)
103. Shi, B., Du, S.S., Su, W.J., Jordan, M.I.: Acceleration via symplectic discretization of high-resolution differential equations. In: NeurIPS, pp. 5744–5752 (2019)

104. Solodov, M.V., Svaiter, B.F.: A hybrid approximate extragradient-proximal point algorithm using the enlargement of a maximal monotone operator. Set Valued Anal. **7**(4), 323–345 (1999)
105. Song, C., Jiang, Y., Ma, Y.: Unified acceleration of high-order algorithms under Hölder continuity and uniform convexity. SIAM J. Optim. (to appear) (2021)
106. Su, W., Boyd, S., Candès, E.J.: A differential equation for modeling Nesterov's accelerated gradient method: theory and insights. J. Mach. Learn. Res. **17**(1), 5312–5354 (2016)
107. Sutherland, W.A.: Introduction to Metric and Topological Spaces. Oxford University Press, Oxford (2009)
108. Tseng, P.: Approximation accuracy, gradient methods, and error bound for structured convex optimization. Math. Program. **125**(2), 263–295 (2010)
109. Vassilis, A., Jean-François, A., Charles, D.: The differential inclusion modeling FISTA algorithm and optimality of convergence rate in the case b $\leq$ 3. SIAM J. Optim. **28**(1), 551–574 (2018)
110. Wibisono, A., Wilson, A.C., Jordan, M.I.: A variational perspective on accelerated methods in optimization. Proc. Natl. Acad. Sci. **113**(47), E7351–E7358 (2016)
111. Wilson, A.C., Mackey, L., Wibisono, A.: Accelerating rescaled gradient descent: fast optimization of smooth functions. In: NeurIPS, pp. 13555–13565 (2019)
112. Wilson, A.C., Recht, B., Jordan, M.I.: A Lyapunov analysis of momentum methods in optimization. J. Mach. Learn. Res. (to appear) (2021)
113. Zhang, J., Mokhtari, A., Sra, S., Jadbabaie, A.: Direct Runge–Kutta discretization achieves acceleration. In: NeurIPS, pp. 3900–3909 (2018)