IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received December 13, 2019, accepted January 6, 2020, date of publication February 3, 2020, date of current version February 17, 2020.

Digital Object Identifier 10.1109/ACCESS.2020.2967064

Implicit Runge-Kutta Methods for Accelerated Unconstrained Convex Optimization

RUIJUAN CHEN AND XIUTING LI
School of Artificial Intelligence and Automation, Huazhong University of Science and Technology, Wuhan 430074, China

Corresponding author: Xiuting Li (xtingli@hust.edu.cn)

This work was supported in part by the National Natural Science Foundation of China under Grant 61903148, and in part by the China Postdoctoral Science Foundation under Grant 2017M622396.

ABSTRACT Accelerated gradient methods have the potential of achieving optimal convergence rates and have successfully been used in many practical applications. Despite this fact, the rationale underlying these accelerated methods remain elusive. In this work, we study gradient-based accelerated optimization methods obtained by directly discretizing a second-order ordinary differential equation (ODE) related to the Bregman Lagrangian. We show that for sufficiently smooth objectives, the acceleration can be achieved by discretizing the proposed ODE using s-stage q-order implicit Runge-Kutta integrators. In particular, we prove that under the assumption of convexity and sufficient smoothness, the sequence of iteration generated by the proposed accelerated method stably converges to the optimal solution at a rate of $O((1-\tilde{C}_{p,q}\cdot\frac{\mu}{L})^N N^{-p})$, where $p \geq 2$ is the parameter in the second-order ODE and $\tilde{C}_{p,q}$ is a constant depending on p and q. Several numerical experiments are given to verify the convergence results.

INDEX TERMS Implicit Runge-Kutta methods, ordinary differential equations, unconstrained convex optimization.

I. INTRODUCTION

Numerous problems in machine learning [1], system identification [2] and optimal control [3]–[5] involves minimizing convex and strongly convex functions. Methods for solving the minimization problems have therefore been extensively developed (cf. [6]). The gradient descent (GD) is one of these methods and it only use gradient information in the optimization procedure so that very large-scale problems can be touched. Currently, gradient-based optimization methods have become the focus of intense research efforts. A central tension in gradient-based optimization methods is the convergence rate. Optimization methods with rapid convergence rates are more popular in practical implementations due to their reasonable computing costs.

Acceleration optimization methods have the potential of achieving faster convergence rates than GD. The heavy-ball (HB) method is an earlier class of acceleration algorithms, which attains the fast convergence rate by incorporating a momentum term into the gradient step [7]. However, it is shown that the HB method hardly ensures global acceleration [7], [8]. In 1983, Nesterov introduced accelerated gradient descent methods that have a global convergence rate [9], [10], showing that an optimal convergence rate is reachable under an oracle model of optimization complexity. From the start, accelerated methods garnered wide attention, which led to many different accelerated methods, such as composite optimization [11], [12], accelerated coordinate descent methods [13], [14] and stochastic optimization [15], [16], to name only a few. More recently, authors of [17] further extended Nesterov's accelerated gradient descent (NAG) to global convex and quasi-strongly convex objectives and obtained linear convergence rates.

The progress of acceleration methods motivates many researchers to explore the rationale underlying the phenomenon of acceleration. Nesterov's original derivation heavily relies on case-specific algebra [18], which is unintuitive and does not easily carry over to general settings. In recent years, a promising direction is to look at acceleration from a continuous-time perspective. Su et al. showed in [19] that the continuous limit of NAG method is a second-order ordinary differential equation (ODE) describing a physical system with vanishing friction. With this ODE and the stability theory of dynamic system at hand, they revalidate

The associate editor coordinating the review of this manuscript and approving it for publication was Fuhui Zhou.

the convergence rate of NAG via the discrete version of a Lyapunov function. Following the limiting arguments, reference [20] derived high-resolution ODEs for the HB method and the NAG method respectively and found the difference of the dynamics corresponding to the two methods. It is also showed that the high-resolution ODEs combing with a general Lyapunov function framework enable the analysis of accelerated convergence rates of NAG. Furthermore, reference [21] explored several discretization schemes for ODEs and found the superiority of the symplectic scheme combing with the high-resolution ODEs in the accelerated rate. Notably, the limit arguments typically require a priori knowledge of existing discrete-time accelerated gradient methods to derive ODEs. Alternatively, Wibisono *et al.* [22] took a variational point of view for the derivation of ODEs. The key point is to use a Lagrangian functional called the Bregman Lagrangian to derive the Euler-Lagrange equations and then elaborately discrete the ODE to generate accelerated methods. It has been shown that the Bregman Lagrangian framework permits a systematic understanding of the acceleration phenomenon among a family of discrete-time accelerated algorithms. Other works in explaining the acceleration phenomenon include unification of mirror and gradient step [23], explicit Runge-Kutta (ExRK) discretization [24]–[26] and the Powerball method [27].

In this paper, we combine the Bregman Lagrangian framework with implicit Runge-Kutta (ImRK) integrator for further analyzing acceleration methods and their connection with continuous dynamics. More specifically, we first leverage the Bregman Lagrangian defined in [22] to derive a second-order ODE without resorting to known accelerated algorithms. Then, we show that the sequence of iterations generated by applying an ImRK integrator to the second-order ODE converges to the optimal solution at an enhanced convergence rate. Additionally, we also theoretically prove the stable convergence of the discrete-time algorithm via an newly designed Lyapunov function. Finally, the effectiveness of the proposed architecture is demonstrated on several convex objectives and its performance is compared with other acceleration methods.

The remainder of the paper is organized as follows. In Section II, we formulate the problem of interest and make some mathematical preliminary on the Bregman Largangian method, the Implicit Runge-Kutta methods and elementary differentials. In Section III, the convergence of the proposed accelerated algorithm is analyzed. In Section IV, we provide some lemmas for our main results. In Section V, numerical results of two convex optimization problems are presented and compared with the performance of gradient descent (GD), NAG and ExRK at different stages. In Section VI, we summarize the work of this paper and discuss future work.

II. PROBLEM FORMULATION AND PRELIMINARY

Consider the following unconstrained optimization problem:

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

where the function $f\in\mathbb{R}^d\rightarrow\mathbb{R}$ is sufficiently smooth and convex. We assume that the optimization problem has an optimal solution, denoted by $x^*\in\mathbb{R}^d$. The aim of this paper is to construct an iterate sequence $\{x_k\}_{k=1}^{\infty}$ such that it converges to x^*. There are numerous strategies for constructing such a sequence. In this paper, we combine the Bregman Lagrangian framework with implicit Runge-Kutta (ImRK) integrator to generate it. For this, we briefly overview the Bregman Lagrangian method as well as the ImRK integrator in the following.

A. THE BREGMAN LAGRANGIAN

Under the ideal scaling assumption, the Bregman Lagrangian can define a variational problem, the solutions to which minimize the objective function f in (1) at an exponential rate. In [22], the Bregman Lagrangian is defined as the following weighted Lagrangian of the mechanical system

$$\mathscr{L}(v,x,t)=e^{\alpha_t+\gamma_t}\big(\mathscr{D}_d(x+e^{-\alpha_t}v,x)-e^{\beta_t}f(x)\big), \tag{2}$$

where the functions $\alpha_t,\beta_t,\gamma_t:T\rightarrow\mathbb{R}$ are arbitrary smooth functions with respect to time, $T\subseteq\mathbb{R}$ is a time interval. $\mathscr{D}_d(y,x)=d(y)-d(x)-\langle\nabla d(x),y-x\rangle$ is the Bregman divergence of distance function $d(\cdot):\mathbb{R}^d\rightarrow\mathbb{R}$. In this paper, we consider the Euclidean setting, i.e., $d(x)=\frac{1}{2}\|x\|^2$, in which case the Bregman divergence reduce to $\mathscr{D}_d(y,x)=\frac{1}{2}\|y-x\|^2$.

According to the calculation of variations, the necessary condition for minimizing the function f is that x is the solution of the following Euler-Lagrange equation,

$$\frac{d}{dt}\left\{\frac{\partial\mathscr{L}}{\partial v}(\dot{x},x,t)\right\}=\frac{\partial\mathscr{L}}{\partial x}\big(\dot{x},x,t\big). \tag{3}$$

For general functions $\alpha_t,\beta_t,\gamma_t$, (3) is actually a second-order differential equation given by

$$\ddot{x}+\big(\dot{\gamma}_t-\dot{\alpha}_t\big)\dot{x}+e^{2\alpha_t+\beta_t}\nabla f(x)=0. \tag{4}$$

When the ideal conditions $\dot{\beta}_t\leq e^{\alpha_t}$ and $\dot{\gamma}_t=e^{\alpha_t}$ is reachable, and we choose the parameters $\alpha_t=\log 2\,p-\log(t+1)$, $\beta_t=p\log(t+1)+\log c$ and $\gamma_t=2\,p\log(t+1)$, where $p,\ c>0$ are constants, then (4) becomes

$$\ddot{x}+\frac{2p+1}{t+1}\dot{x}+4cp^2(t+1)^{p-2}\nabla f(x)=0. \tag{5}$$

Moreover, let $y=[v;x]\in\mathbb{R}^{2d}$, $v=\dot{x}$, then (5) can be written as a dynamical system as follows,

$$\dot{y}=\left[-\frac{2p+1}{t+1}v-p^2(t+1)^{p-2}\nabla f(x);\ v\right]:=F(y), \tag{6}$$

where $c=1/4$. Denote by $F_v=-\frac{2p+1}{t+1}v-p^2(t+1)^{p-2}\nabla f(x)$ and $F_x=v$ the component of F. We have $F=[F_v,F_x]$.

B. IMPLICIT RUNGE-KUTTA INTEGRATORS

Runge-Kutta methods offer a powerful class of numerical integrators, encompassing several basic discretization schemes. In this subsection, we briefly recap implicit Runge-Kutta (ImRK) integrators.

Definition 1: Given a dynamical system $\dot{y} = F(y)$. Let $y_0 := [\mathbf{0}; x_0]$ be the current point and h be the step size. An s-stage Runge-Kutta method generates the next step via the updating procedure,

$$z_i = y_0 + h\sum_{j=1}^{s} a_{ij}F\big(z_j\big), \quad i = 1, 2, \ldots, s, \tag{7}$$

$$\Phi_h(y_0) = y_0 + h\sum_{i=1}^{s} b_i F\big(z_i\big), \tag{8}$$

where x_0 is an arbitrary point in $\mathbb{R}^d$, $\mathbf{0} \in \mathbb{R}^d$ with all components being 0, a_{ij} and b_i are suitable coefficients defined by the integrator. $\Phi_h(y_0)$ is the estimation of the state y after time step size h, while z_i, $i = 1, \ldots, s$ is a few neighboring points where the value of $F(z_i)$ is evaluated.

In general, by combining the gradients of multiple evaluation points and matching the Taylor expansion coefficients, the higher accuracy can be obtained. Based on this, we first introduce the definition of integrator order.

Definition 2: ([28]) Let $\varphi_h(y_0)$ be the true solution to the ODEs (6) with initial condition y_0. We say that an integrator $\Phi_h(y_0)$ has order q if its discretization error shrinks as

$$\big\|\Phi_h(y_0) - \varphi_h(y_0)\big\| = O\big(h^{q+1}\big), \quad \text{as } h \to 0. \tag{9}$$

Definition 3: ([29]) The function $R(z)$ is called the stability function of the method. It can be interpreted as the numerical solution after one step for the famous Dahlquist test equation,

$$\dot{u} = \lambda u, \quad u_0 = 1, \ w = h\lambda. \tag{10}$$

The set

$$S = \{w \in \mathbb{C}, |R(w)| \leq 1\}, \tag{11}$$

is called the stability domain of the method.

Proposition 4: ([29]) The s-stage implicit Runge-Kutta method (7)-(8) applied to $\dot{u} = \lambda u$ yields $\Phi_h(y_0) = R(h\lambda)y_0$ with

$$R(w) = 1 + wb^T(I - wA)^{-1}\mathbf{1}, \tag{12}$$

where $b^T = (b_1, \ldots, b_s)$, $A = (a_{ij})_{i,j=1}^{s}$ and $\mathbf{1} = (1, \ldots, 1)^T$.

Definition 5: ([30], [31]) A method is called A-stable, if its stability domain satisfies

$$S \supset \mathbb{C}^{-1} = \{w, \ \mathrm{Re}\, w \leq 0\}. \tag{13}$$

Lemma 6: ([29]) A s-stage Runge-Kutta method (7)-(8) is A-stable if and only if R(z) is analytic on $\mathbb{C}^{-1}$, and

$$|R(iz)| < 1, \quad \forall z \in \mathbb{R}, \tag{14}$$

where i is the imaginary unit.

Throughout this paper, we make the following assumption for the implicit Runge-Kutta method.

Assumption 7: The implicit Runge-Kutta method (7)-(8) is A-stable.

C. ELEMENTARY DIFFERENTIALS

In this subsection, we recall some facts on elementary differentials. Unless otherwise specified, all the results presented in this subsection have been proved in [32]. Given a dynamical system $\dot{y} = F(y)$, we want to find a convenient way to express and compute its higher order derivatives. For this, let τ denote a tree structure, and $|\tau|$ is the number of nodes in τ.

Definition 8: The set of (rooted) trees τ is recursively defined as follows: a) the graph $\bullet$ with only one vertex (called the root) belongs to τ; b) if $\tau_1, \ldots, \tau_m \in \tau$, then the graph obtained by grafting the roots of $\tau_1, \ldots, \tau_m$ to a new vertex also belongs to τ. It is denoted by

$$\tau = [\tau_1, \ldots, \tau_m], \tag{15}$$

and the new vertex is the root of τ.

Definition 9: For a tree τ, the elementary differential is a mapping $F(\tau) : \mathbb{R}^d \to \mathbb{R}^d$, defined recursively by $F(\bullet)(y) = F(y)$ and

$$F(\tau)(y) = \nabla^m F(y)\big[F(\tau_1)(y), \ldots, F(\tau_m)(y)\big], \tag{16}$$

for $\tau = [\tau_1, \ldots, \tau_m]$ and $\sum_{i=1}^{m} |\tau_i| = |\tau| - 1$.

With these definitions, the following results hold and its proof is obtained by recursively applying the product rule [32].

Lemma 10: The q-th order derivative of the exact solution to $\dot{y} = F(y)$ is given by

$$y^{(q)}(t_0) = F^{(q-1)}(y_0) = \sum_{|\tau|=q} \alpha(\tau)F(\tau)(y_0), \tag{17}$$

where $y(t_0) = y_0$, $\alpha(\tau)$ is a positive integer determined by τ and the number of occurrences of the tree pattern τ.

The expression for $\frac{d^qF(z_i)}{dh^q}$ can be calculated in the same way as in Lemma 10 by the Leibniz rule.

Lemma 11: For a Runge-Kutta method defined in Definition 1, if F is q-th differentiable, then

$$\frac{d^q\Phi_h(y_0)}{dh^q} = \sum_{i\leq s} b_i\left[h\frac{d^qF(z_i)}{dh^q} + q\frac{d^{q-1}F(z_i)}{dh^q}\right], \tag{18}$$

where $\frac{d^qF(z_i)}{dh^q}$ has the same structure as $F^{(q)}(y)$ in Lemma 10, except that we need to replace all F in the expression by $\frac{dz_i}{dh}$ and all $\nabla^n F(y)$ by $\nabla^n F(z_i)$.

III. CONVERGENCE ANALYSIS

In this section, we combine the Bregman Lagrangian framework with implicit Runge-Kutta (ImRK) integrator to derive our accelerated optimization algorithm and then analyze the convergence of the obtained optimization algorithm. Reference [22] pointed out that simple discretization (such as Euler method) applied to ODEs is difficult to guarantee a stable discrete-time algorithm . Based on this, we propose to use the implicit Runge-Kutta integrators to discrete the second-order ODE defined in (6) and then design the stable discrete-time algorithm. Our designed optimization algorithm is summarized in Algorithm 1.

In the following, we first give the definition of the stable discrete-time algorithm.

Algorithm 1 Computing of $\{x_k\}$

```
Constants $L$ is the same as in Assumption 13
Set the initial state $y_0 = [\mathbf{0}; x_0] \in \mathbb{R}^{2d}$ with arbitrary initial time $t_0 \geq 0$ and $\mathscr{E}_0 := \mathscr{E}(y_0)$
Choosing step size $h = \frac{1}{4}\left(\frac{\mathscr{E}_0^2}{C_p(q+1)^2}\right)^{1/q}\frac{1}{(\mathscr{E}_0+L+1)^{1+2/q}}$ where $C_p = c_1^p(p+1)!$ and $c_1 > 1$ is a constant
$x_k \leftarrow$ $s$-stage $q$-order implicit Runge-Kutta integrator $(F, y_0, k, h)$, where $F$ is defined in equation (6)
Return $x_k$
```

Definition 12: Assume that a discrete-time algorithm is obtained by applying the A-stable discrete scheme (7)-(8) to the ODE (6). We call the discrete-time algorithm stable, if the discrete-time algorithm retains the convergence rate of the underlying ODE (6).

In order to take advantage of the order conditions of the implicit Runge-Kutta integrators, we make the following assumptions about the boundedness of higher-order derivatives of the function f.

Assumption 13: Based on (6), we assume $\mathscr{F} \subseteq C_L^{k,q}(\mathbb{R}^d)$ with $k, q \geq p$, which means that for any $f \in \mathscr{F}$, it is k times continuous differentiable on $\mathbb{R}^d$ and its q-th derivative is Lipschitz continuous on $\mathbb{R}^d$ with a constant $L \geq 1$,

$$\left\|f^{(q)}(x) - f^{(q)}(y)\right\| \leq L\|x - y\|,$$

for all $x,\ y \in \mathbb{R}^d$. Moreover, assume that $\mathscr{F}$ is compact, then for $i = 1, \ldots, q+1$,

$$\left\|f^{(i)}(x)\right\| \leq L. \tag{19}$$

Assumption 14: Function $f(x)$ is μ-strongly convex, i.e.,

$$f(y) \geq f(x) + \langle \nabla f(x), y - x\rangle + \frac{\mu}{2}\|y - x\|^2, \tag{20}$$

for all $x,\ y \in \mathbb{R}^d$.

Lemma 15: Let $f(x)$ be twice differentiable on $\mathbb{R}^d$ and $x^* := \arg\min_x f(x)$. Then

$$\frac{1}{4}\lambda_1\|x - x^*\|^2 \leq f(x) - f(x^*), \tag{21}$$

where λ_1 is the smallest eigenvalue of matrix $f^{(2)}(x^*)$.

Proof: See the proof of Theorem 1.2.3 in [10]. ■

Remark 16: Based on Assumption 14 and Lemma 15, one can prove that the second derivative of f satisfies $f^{(2)}(x) \geq \ell I_d$, where I_d is the identity matrix and $\ell = \min\{\lambda_1, \mu, 1\}$.

Lyapunov functions play a central role in the convergence analysis of ODEs in both continuous time and discrete time. For the nonlinear dynamical system (6), we define the Lyapunov function as follows,

$$\begin{aligned}\mathscr{E}(y(t)) &= \mathscr{E}\big([v; x], t\big)\\ &= 2(t+1)^p\big(f(x) - f(x^*)\big) + \frac{(t+1)^2}{4p^2}\|v\|^2\\ &\quad + 3\left\|x + \frac{t+1}{2p}v - x^*\right\|^2.\end{aligned} \tag{22}$$

Remark 17: The Lyapunov function is different from ones in [25] and [22] in terms of the coefficients and the quadratic term. The difference are the key to prove the improved convergence of the direct discrete optimization algorithm proposed in this paper.

Now, we focus on obtaining the stability of the continuous system (6). We start by introducing some notations. Given a vector $y_0 = [\mathbf{0}; x_0] \in \mathbb{R}^{2d}$ with arbitrary initial time $t_0 \geq 0$, we define the neighborhood of y_0 as $U_\delta(y_0) := \{y = [v; x] |\ \|x - x_0\| \leq \delta,\ \|v - v_0\| \leq \delta\}$ and time interval $|t - t_0| \leq \sigma$ where $0 < \sigma < 1$, $t_0 \geq 0$ and $\delta := 1/(t_0 + 1)$. Let $\mathscr{E}(y_0) := \mathscr{E}_0$, without loss of generality, assume that $\mathscr{E}_0 \geq 1$. The stability of the continuous system is justified by the following Lemma.

Lemma 18: Consider $y = [v; x] \in \mathbb{R}^{2d}$ as a trajectory of the dynamical system (6). Let the Lyapunov function $\mathscr{E}$ be defined by (22). Then, for any trajectory y, the time derivative $\dot{\mathscr{E}}(y)$ is non-positive and bounded above, more precisely,

$$\dot{\mathscr{E}}(y) \leq -\frac{p}{C'(t+1)}\mathscr{E}(y), \tag{23}$$

where $C' > 0$ is a constant.

Proof: According to the dynamical system (6), we can write

$$\dot{x} = v, \quad \ddot{x} = \dot{v} = -\frac{2p+1}{t+1}v - p^2(t+1)^{p-2}\nabla f(x). \tag{24}$$

Then, it can be proved that

$$\begin{aligned}\dot{\mathscr{E}} &= \frac{(t+1)^2}{4p^2}\big\langle 2v,\ \dot{v}\big\rangle + \frac{2(t+1)}{4p^2}\big\langle v,\ v\big\rangle\\ &\quad + 2(t+1)^p\big\langle \nabla f(x),\ \dot{x}\big\rangle + 2p(t+1)^{p-1}\big(f(x) - f(x^*)\big)\\ &\quad + 2\cdot 3\Big\langle x + \frac{t+1}{2p}v - x^*,\ \dot{x} + \frac{\dot{x}}{2p} + \frac{t+1}{2p}\ddot{x}\Big\rangle\\ &= \frac{2(t+1)^2}{4p^2}\Big\langle \dot{x},\ \ddot{x} + \frac{2p+1}{(t+1)}\dot{x}\Big\rangle - \frac{2(t+1)}{4p^2}\big\langle \dot{x},\ 2p\dot{x}\big\rangle\\ &\quad + 2(t+1)^p\big\langle \nabla f(x),\ \dot{x}\big\rangle + 2\,p(t+1)^{p-1}\big(f(x) - f(x^*)\big)\\ &\quad + 2\frac{3(t+1)}{2p}\Big\langle x + \frac{t+1}{2p}\dot{x} - x^*,\ \ddot{x} + \frac{2p+1}{(t+1)}\dot{x}\Big\rangle\\ &= \frac{(t+1)^2}{2p^2}\big\langle \dot{x},\ -p^2(t+1)^{p-2}\nabla f(x)\big\rangle - \frac{(t+1)}{p}\|\dot{x}\|^2\\ &\quad + 2(t+1)^p\big\langle \nabla f(x),\ \dot{x}\big\rangle + 2\,p(t+1)^{p-1}\big(f(x) - f(x^*)\big)\\ &\quad - \frac{3(t+1)}{p}\Big\langle x + \frac{t+1}{2p}\dot{x} - x^*,\ p^2(t+1)^{p-2}\nabla f(x)\Big\rangle.\end{aligned}$$

By introducing the term $\frac{p\ell}{16(t+1)}\left\|x + \frac{t+1}{2p}\dot{x} - x^*\right\|^2$, we have

$$\begin{aligned}\dot{\mathscr{E}} &= -\frac{(t+1)^p}{2}\big\langle \dot{x},\ \nabla f(x)\big\rangle - \frac{(t+1)}{p}\|\dot{x}\|^2\\ &\quad + 2(t+1)^p\big\langle \nabla f(x),\ \dot{x}\big\rangle + 2\,p(t+1)^{p-1}\big(f(x) - f(x^*)\big)\\ &\quad - \frac{3(t+1)^2}{2\,p^2}\big\langle \dot{x},\ p^2(t+1)^{p-2}\nabla f(x)\big\rangle - \frac{3(t+1)}{p}\big\langle x - x^*,\end{aligned}$$

$$p^2(t+1)^{p-2}\nabla f(x)\big\rangle - \frac{p\ell}{16(t+1)}\left\|x+\frac{t+1}{2p}\dot{x}-x^*\right\|^2 + \frac{p\ell}{16(t+1)}\left\|x+\frac{t+1}{2p}\dot{x}-x^*\right\|^2 + \frac{1}{2}p(t+1)^{p-1}\big(f(x) - f(x^*)\big) - \frac{1}{2}p(t+1)^{p-1}\big(f(x)-f(x^*)\big).$$

Following the elementary inequality $(a+b)^2 \leq 2\,a^2 + 2\,b^2$ with any real numbers a and b, one can prove

$$\begin{aligned}\dot{\mathscr{E}} \leq& -\frac{(t+1)^p}{2}\big\langle \dot{x},\ \nabla f(x)\big\rangle - \frac{3}{2}(t+1)^p\big\langle \dot{x},\ \nabla f(x)\big\rangle \\ &+ 2(t+1)^p\big\langle \nabla f(x),\ \dot{x}\big\rangle - \frac{(t+1)}{p}\|\dot{x}\|^2 - \frac{1}{2}p(t+1)^{p-1} \\ &\cdot\big(f(x)-f(x^*)\big) - \frac{p\ell}{16(t+1)}\left\|x+\frac{t+1}{2p}\dot{x}-x^*\right\|^2 \\ &+ 2\,p(t+1)^{p-1}\big(f(x)-f(x^*)\big) - 3\,p(t+1)^{p-1}\big\langle x-x^*, \\ &\quad\nabla f(x)\big\rangle + \frac{1}{2}p(t+1)^{p-1}\big(f(x)-f(x^*)\big) \\ &+ \frac{2p\ell}{16(t+1)}\|x-x^*\|^2 + \frac{2p\ell}{16(t+1)}\cdot\frac{(t+1)^2}{4p^2}\|\dot{x}\|^2 \\ =& -\frac{(t+1)}{p}\|\dot{x}\|^2 + \frac{(t+1)\ell}{32\,p}\|\dot{x}\|^2 - \frac{1}{2}p(t+1)^{p-1} \\ &\cdot\big(f(x)-f(x^*)\big) - \frac{p\ell}{16(t+1)}\left\|x+\frac{t+1}{2p}\dot{x}-x^*\right\|^2 \\ &- 3\,p(t+1)^{p-1}\big\langle x-x^*,\ \nabla f(x)\big\rangle + \frac{5}{2}p(t+1)^{p-1} \\ &\big(f(x)-f(x^*)\big) + \frac{2\,p\ell}{16(t+1)}\|x-x^*\|^2.\end{aligned}$$

From Assumption 15, we have $\frac{\ell}{16(t+1)}\|x - x^*\|^2 \leq \frac{1}{4}(f(x)-f(x^*)) \leq \frac{1}{4}(t+1)^p(f(x)-f(x^*))$ and then

$$\begin{aligned}\dot{\mathscr{E}} \leq& -\frac{t+1}{2p}\|\dot{x}\|^2 - \frac{1}{2}p(t+1)^{p-1}\big(f(x)-f(x^*)\big) \\ &- \frac{p\ell}{16(t+1)}\left\|x+\frac{t+1}{2p}\dot{x}-x^*\right\|^2 - 3p(t+1)^{p-1}\big\langle x-x^*, \\ &\nabla f(x)\big\rangle + \frac{5}{2}p(t+1)^{p-1}\big(f(x)-f(x^*)\big) \\ &+ \frac{1}{2}p(t+1)^{p-1}\big(f(x)-f(x^*)\big) \\ =& -\frac{t+1}{2p}\|\dot{x}\|^2 - \frac{1}{2}p(t+1)^{p-1}\big(f(x)-f(x^*)\big) \\ &- \frac{p\ell}{16(t+1)}\left\|x+\frac{t+1}{2p}\dot{x}-x^*\right\|^2 - 3p(t+1)^{p-1}\big\langle x-x^*, \\ &\nabla f(x)\big\rangle + 3p(t+1)^{p-1}\big(f(x)-f(x^*)\big). \qquad (25)\end{aligned}$$

From the convexity of f, then

$$\begin{aligned}\dot{\mathscr{E}} \leq& -\frac{t+1}{2p}\|\dot{x}\|^2 - \frac{1}{2}p(t+1)^{p-1}\big(f(x)-f(x^*)\big) \\ &- \frac{p\ell}{16(t+1)}\left\|x+\frac{t+1}{2p}\dot{x}-x^*\right\|^2 \\ \leq& -\frac{(t+1)\ell}{192\,p}\|\dot{x}\|^2 - \frac{\ell}{24}p(t+1)^{p-1}\big(f(x)-f(x^*)\big) \\ &- \frac{\ell}{16}\frac{p}{t+1}\left\|x+\frac{t+1}{2p}v-x^*\right\|^2 \\ =& -\frac{p\ell}{48(t+1)}\mathscr{E}(y) \\ =& -\frac{p}{C'(t+1)}\mathscr{E}(y), \qquad (26)\end{aligned}$$

where $C' = 48/\ell > 0$ is a constant. ■

Before giving the main convergence results in this paper, we first present some properties enjoyed by the Lyapunov function (22).

Proposition 19: Under the Assumption 13, we discretize (5) with a s-stage q-order Runge-Kutta integrator. By setting

$$h \leq \left(\frac{1}{4C'}\frac{\mathscr{E}_0}{C_p(t_0+1)^p(q+1)^2}\right)^{1/q}\frac{1}{(\mathscr{E}_0+L+1)^{1+2/q}},$$

where $C_p = c_1^p(p+1)!$ is a constant, then

$$\mathscr{E}(y_N) \leq \left(1 - \frac{hp}{2C'(t_0+1)}\right)^N \mathscr{E}_0^2. \qquad (27)$$

Proof: By Taylor's theorem and triangle inequality, we know that

$$\begin{aligned}&\big|\mathscr{E}(\Phi_h(y_0)) - \mathscr{E}(\varphi_h(y_0))\big| \\ &\quad\leq h^{q+1}\max_{0\leq\lambda\leq h}\left(\left|\frac{d^{q+1}}{dh^{q+1}}\mathscr{E}(\Phi_\lambda(y_0))\right| + \left|\frac{d^{q+1}}{dh^{q+1}}\mathscr{E}(\varphi_\lambda(y_0))\right|\right).\end{aligned}$$

Since $\mathscr{E}(\varphi_h(y_0)) \leq \left(1 - \frac{hp}{C'(t_0+1)}\right)\mathscr{E}_0$, then

$$\begin{aligned}&\mathscr{E}(\Phi_h(y_0)) \\ &\leq \left(1-\frac{hp}{C'(t_0+1)}\right)\mathscr{E}_0 + h^{q+1} \\ &\quad\cdot\max_{0\leq\lambda\leq h}\left(\left|\frac{d^{q+1}}{dh^{q+1}}\mathscr{E}(\Phi_\lambda(y_0))\right| + \left|\frac{d^{q+1}}{dh^{q+1}}\mathscr{E}(\varphi_\lambda(y_0))\right|\right) \\ &\leq \left(1-\frac{hp}{C'(t_0+1)}\right)\mathscr{E}_0 + h^{q+1}\left(\frac{C_{p,q+1}}{(t_0+1)}(1+L+\mathscr{E}_0)^{q+2}\right. \\ &\quad\left.+\frac{C'_{p,q+1}}{(t_0+1)}(1+L+\mathscr{E}_0)^{q+2}\right) \\ &\leq \left(1-\frac{hp}{C'(t_0+1)}\right)\mathscr{E}_0^2 + h\cdot h^q\frac{2\widehat{C}_{p,q+1}}{t_0+1}(1+L+\mathscr{E}_0)^{q+2}.\end{aligned}$$

The second inequality follows by Lemma 24 and Lemma 25 in Section IV. The last inequality obtained from $\widehat{C}_{p,q+1} := \max\{C_{p,q+1}, C'_{p,q+1}\}$. Without loss of generality, assume that $C_{p,q+1} \geq C'_{p,q+1}$, i.e., $\widehat{C}_{p,q+1} = C_{p,q+1} = C_p(t_0+1)^p(q+1)^2$.

Choosing step-size h to satisfy

$$h \leq \left(\frac{1}{4C'}\frac{\mathscr{E}_0^2}{C_p(t_0+1)^p(q+1)^2}\right)^{1/q}\frac{1}{(\mathscr{E}_0+L+1)^{1+2/q}}.$$

Then we have

$$\mathscr{E}(\Phi_h(y_0)) \leq \left(1-\frac{hp}{2C'(t_0+1)}\right)\mathscr{E}_0^2.$$

The last inequality follows by the choice of h. ■

Remark 20: It has been shown in [28] that the implicit Runge-Kutta integrator with the same stage has higher convergence order than the corresponding explicit methods.
In the following, we use Proposition 19 to obtain the stable convergence of our proposed algorithm.

Theorem 21: Consider (6), suppose that f satisfies Assumption 13 and Assumption 14. Further, let q be the order of the Runge-Kutta integrator with s-stage used in (7)-(8), N be the total number of iterations and x_0 be the initial point. It there exists constants $\mathscr{E}_0 = L$ and C_p such that the step size $h = \frac{1}{4}\left(\frac{\mathscr{E}_0^2}{C_p(q+1)^2}\right)^{1/q} \frac{1}{(\mathscr{E}_0+L+1)^{1+2/q}}$, then the iterate x_N generated after N times satisfies the following inequality,

$$f(x_N) - f(x^*) \leq \tilde{C}\left(1 - \tilde{C}_{p,q} \cdot \frac{\mu}{L}\right)^N N^{-p}, \tag{28}$$

where $\tilde{C} = 4^p \mathscr{E}_0^2 \left(\frac{C_p(q+1)^2}{\mathscr{E}_0}\right)^{p/q} (\mathscr{E}_0 + L + 1)^{2p}$ and $\tilde{C}_{p,q} = \frac{p}{3600(t_0+1)}\left(\frac{1}{C_p(q+1)^2}\right)^{1/q}$.

Proof: According to Definition 12, we first prove the convergence of the algorithm according to the A-stability of the discrete format. When the objective function f is quadratic, denoted as $f(x) = 1/2\, x^T Ux + Vx + W$, where U is a positive definite, symmetric $n \times n$ matrix. Then (6) is a linear dynamical system

$$\dot{y} = F(y) = \Theta y + \Lambda, \tag{29}$$

where $\Theta = \begin{bmatrix} -\frac{2p+1}{t+1} I & p^2(t+1)^{p-2} U \\ I & 0 \end{bmatrix}$ and $\Lambda = \begin{bmatrix} V \\ 0 \end{bmatrix}$.
Using ImRK (7)-(8) to discretize (29), the following algorithm can be obtained

$$z_{n,i} = y_n + h \sum_{j=1}^{s} a_{ij} \mathbf{F}\big(z_{n,j}\big), \quad i = 1, 2, \ldots, s, \tag{30}$$

$$y_{n+1} = y_n + h \sum_{i=1}^{s} b_i \mathbf{F}\big(z_{n,i}\big), \tag{31}$$

We rewrite (30)-(31) as matrix format

$$\begin{aligned} z_n &= \mathbf{1} \otimes y_n + h(A \otimes I)\mathbf{F}\big(z_n\big) \\ &= \mathbf{1} \otimes y_n + h(A^T \otimes I)(\Theta z_n + \Lambda), \qquad (32) \\ y_{n+1} &= y_n + h(b^T \otimes I)\mathbf{F}\big(z_n\big), \qquad (33) \end{aligned}$$

where $\mathbf{1} = [1, \cdots, 1]^T \in \mathbb{R}^s$. Then we have $z_n = (I - h(A \otimes I)(\mathbf{1} \otimes \Theta))^{-1}\mathbf{1} y_n + (I - h(A \otimes I)(\mathbf{1} \otimes \Theta))^{-1} h(A \otimes I)(\mathbf{1} \otimes \Lambda)$. Substituting it into (33), we have

$$\begin{aligned} &y_{n+1} \\ &= y_n + h(b^T \otimes I)(\mathbf{1} \otimes \Theta)(I - h(A \otimes I)(\mathbf{1} \otimes \Theta))^{-1}(\mathbf{1} \otimes y_n) \\ &\quad + h(b^T \otimes I)\Theta(I - h(A \otimes I)(\mathbf{1} \otimes \Theta))^{-1} h(A \otimes I)(\mathbf{1} \otimes \Lambda) \\ &\quad + (\mathbf{1} \otimes \Lambda), \qquad (34) \end{aligned}$$

which means the proposed optimization algorithm has a linear format,

$$y_{n+1} = \Gamma y_n + \Upsilon, \tag{35}$$

where $\Gamma = I + h\big((b^T \otimes I)(\mathbf{1} \otimes \Theta)\big)(I - h(A \otimes I)(\mathbf{1} \otimes \Theta))^{-1}(\mathbf{1} \otimes I)$. Because of the A-stability of the implicit Runge-Kutta method, the spectral norm $\rho(\Gamma) < 1$ can be obtained according to the definition of A-stability, which means the convergence of the algorithm (35). When f is a nonlinear function, one can do a similar analysis.

Next, we continue to analyze the convergence rate of the proposed algorithm. If the step size h satisfies the condition in Proposition 19, then from the definition of Lyapunov function (22), we can see that

$$f(x_N) - f(x^*) \leq \frac{\mathscr{E}(y_N)}{t_N^p} \leq \left(1 - \frac{hp}{2C'(t_0+1)}\right)^N \frac{1}{(1+Nh)^p} \mathscr{E}_0.$$

If we choose the step size as $h = \frac{1}{4}\left(\frac{\mathscr{E}_0^2}{C_p(q+1)^2}\right)^{1/q} \frac{1}{(\mathscr{E}_0+L+1)^{1+2/q}}$. It is easy to prove that h satisfies

$$h \leq \left(\frac{1}{4C'} \frac{\mathscr{E}_0^2}{C_p(t_0+1)^p(q+1)^2}\right)^{1/q} \frac{1}{(\mathscr{E}_0 + L + 1)^{1+2/q}}.$$

Then we have

$$\begin{aligned} &f(x_N) - f(x^*) \\ &\leq \left(1 - \frac{p}{2C'(t_0+1)} \frac{1}{4}\left(\frac{\mathscr{E}_0^2}{C_p(q+1)^2}\right)^{1/q} \frac{1}{(\mathscr{E}_0 + L + 1)^{1+2/q}}\right)^N \\ &\quad \cdot N^{-p} \cdot 4^p \left(\frac{C_p(q+1)^2}{\mathscr{E}_0^2}\right)^{p/q} (\mathscr{E}_0 + L + 1)^{2p} \cdot \mathscr{E}_0^2, \qquad (36) \end{aligned}$$

which means that

$$f(x_N) - f(x^*) \leq \tilde{C}(1 - C)^N N^{-p}, \tag{37}$$

where $\tilde{C} = 4^p \mathscr{E}_0^2 \left(\frac{C_p(q+1)^2}{\mathscr{E}_0^2}\right)^{p/q} (\mathscr{E}_0 + L + 1)^{2p}$ and $C = \frac{p}{8\ C'(t_0+1)} \frac{1}{(\mathscr{E}_0+L+1)^{1+2/q}} \left(\frac{\mathscr{E}_0^2}{C_p(q+1)^2}\right)^{1/q}$ are constants depending on p and q.

Furthermore, if we choose $\mathscr{E}_0 = L$, then C satisfies

$$\begin{aligned} C &\geq \frac{\mu p}{8 \cdot 48(t_0+1)} \frac{1}{(2L+1)^{1+2/q}} \left(\frac{L^2}{C_p'(q+1)^2}\right)^{1/q} \\ &= \frac{p}{8 \cdot 48(t_0+1)} \left(\frac{1}{C_p(q+1)^2}\right)^{1/q} \cdot \frac{\mu}{2L+1} \left(\frac{L}{2L+1}\right)^{2/q} \\ &\geq \tilde{C}_{p,q} \cdot \frac{\mu}{L}, \qquad (38) \end{aligned}$$

where the last inequality is obtained from $L \geq 1$ and $\tilde{C}_{p,q} = \frac{p}{3600(t_0+1)}\left(\frac{1}{C_p(q+1)^2}\right)^{1/q}$ is determined by p and q. Then from (37), we have

$$f(x_N) - f(x^*) \leq \tilde{C}\left(1 - \tilde{C}_{p,q} \cdot \frac{\mu}{L}\right)^N N^{-p}. \tag{39}$$

■

It is noted from (28) that when the number of iterations N tends to infinity, $(1 - \tilde{C}_{p,q}\frac{\mu}{L})^N$ is a higher order infinity smaller than N^{-p}, indicating that the final convergence rate of the algorithm is mainly determined by $(1 - \tilde{C}_{p,q}\frac{\mu}{L})^N$. According to the conclusion of convergence in Theorem 21, it can be noted that when p is fixed, h is positively correlated with q and with the increase of q, the step size h can be taken in a larger range. When q is fixed, h is negatively correlated with p, and with the increase of p, the step size h can only be taken in a smaller range. Additionally, according to the definition of $\tilde{C}_{p,q}$ in (28), $\tilde{C}_{p,q}$ is positively correlated with q, and the increase of q can accelerate the convergence of the algorithm. On the contrary, $\tilde{C}_{p,q}$ is negatively correlated with p, and the increase of p will reduce the convergence rate of the algorithm.

IV. BOUNDEDNESS OF DERIVATIVES

In this section, we present several key lemmas to prove Proposition 19. First, the bounded properties of higher derivatives of $\varphi_h(y_0)$ and $\Phi_h(y_0)$ is given in the following lemma.

Lemma 22: Given the state $y = [v; x]$. If Assumption 13 holds, then for $n = 1, \dots, q+1$, we have

$$\left\| \frac{d^n \varphi_h(y_0)}{dh^n} \right\| \leq C_0(1 + L + \mathscr{E}_0)^n, \tag{40}$$

and

$$\left\| \frac{d^n \Phi_h(y_0)}{dh^n} \right\| \leq C_1(1+L+\mathscr{E}_0)^n + C_1' h(1+L+\mathscr{E}_0)^{n-1}, \tag{41}$$

where the constants C_0, C_1 and C_1' are determined by p, q and the integrator.

Proof: Notice that the system dynamic $F : \mathbb{R}^{2d} \to \mathbb{R}^{2d}$ in (12) is a vector valued multivariate function. We denote its i-th order derivatives by $\nabla^i F(y)$, which is a $(2d)_1 \times \cdots \times (2d)_{i+1}$ tensor. The tensor is symmetric by the continuity and the Schwartz theorem. As a shorthand, we use $\nabla^i F$ to denote $\nabla^i F(y)$. We know that $y^{(i)} = F^{(i-1)}(y) = \frac{d^i y}{dt^i}$. Notice that $F^{(i-1)}(y)$ is a vector, that is,

$$\begin{aligned} y^{(1)} &= F, \quad y^{(2)} = F^{(1)} = \nabla F(F), \\ y^{(3)} &= F^{(2)} = \nabla^2 F(F, F) + \nabla F(\nabla F(F)). \end{aligned}$$

The derivative $\nabla^i F(y)$ can be interpreted as a linear map $\nabla^i F : \mathbb{R}^{2d} \times \cdots \times \mathbb{R}^{2d} \to \mathbb{R}^{2d}$. $\nabla^2 F(F_1, F_2)$ is a mapping from F_1 and F_2 to an element in European space $\mathbb{R}^{2d}$. Since enumerating the expressions is tedious, we aim to compactly express them with elementary differentials summarized in Section II-C (see Chapter 3.1 in [32] for more details). First we bound $\nabla^i F$ by explicitly computing its entries. Let $a(t) = -p^2(t+1)^{p-2}$ and $b(t) = -\frac{2p+1}{t+1}$. Based on the definition in (12), we have

$$\begin{aligned} &\frac{\partial^i F}{\partial x^i} = \left[a(t)\nabla^{i+1} f(x); \mathbf{0} \right], \quad \frac{\partial F}{\partial v} = [b(t)\mathbf{1}; \mathbf{1}], \\ &\frac{\partial^{i+j} F}{\partial v^j \partial x^i} = \mathbf{0}, \quad \frac{\partial^j F}{\partial v^j} = \mathbf{0}, \ j \geq 2, \end{aligned}$$

where $\mathbf{1}, \mathbf{0} \in \mathbb{R}^d$ are vectors with all components are 1 and 0, respectively. For any vector $y = [v; x]$, we can show that the norm of $\nabla^n F$ is upper bounded by

$$\left\| \nabla^n F(F_1, \dots, F_n) \right\| \leq \left\| a(t) \nabla^{n+1} f(x) \right\| \cdot \Pi_{i=1}^n \left\| F_{i2} \right\|, \tag{42}$$

where $[n] = \{1, 2, \dots, n\}$, $\Lambda \subset [n]$ are the index sets and $F_i = [F_{i1}; F_{i2}]$, $F_{i1}, \ F_{i2} \in \mathbb{R}^d$, $i = 1, 2, \dots, n$. Finally, we are ready to bound the time derivative. We first bound the elementary differential $f(\tau)$ defined in Definition 9. Let $F(\tau) = F(\tau)(y)$ for convenience. Using the definition of the Lyapunov function $\mathscr{E}$, it can be shown that

$$\|v\| \leq \|v_0\| + \|v - v_0\| \leq \frac{2p\mathscr{E}_0^{1/2}}{t_0 + 1} + \delta \leq \frac{p+1}{t_0+1}(\mathscr{E}_0 + 1). \tag{43}$$

Substituting (43) and (19) into (42), then

$$\left\| \nabla^n F(F_1, \dots, F_n) \right\| \leq c_1 L (t_0 + 1)^{p-2} \cdot \Pi_{i=1}^n \|F_{i2}\|,$$

where c_0 depend on n and p.

Let $|\tau| = n$ and it has m subtrees attached to the root, i.e., $\tau = [\tau_1, \dots, \tau_m]$ with $\sum_{i=1}^m |\tau_i| = n - 1$. Then we have

$$\left\| \nabla^m F(F(\tau_1), \dots, F(\tau_m)) \right\| \leq c_0 L (t_0 + 1)^{p-2} \Pi_{i=1}^m \|F_x(\tau_i)\|.$$

We know that $\|F_x(\tau)\| \leq \|v\|$. Based on $t_0 - \sigma \leq t \leq \sigma + t_0$, there always exists some constants $c_1, c_2 > 0$ such that $c_2(t_0+1) \leq t + 1 \leq c_1(t_0 + 1)$ and

$$\begin{aligned} \|F_v(t)\| &\leq \frac{2p+1}{t+1}\|v\| + p^2(t+1)^{p-2}\|\nabla f(x)\| \\ &\leq \frac{2p+1}{c_2(t_0+1)} \frac{1+\mathscr{E}_0}{t_0+1} + p^2 c_1^{p-2}(t_0+1)^{p-2} L \\ &\leq c_p \frac{1 + \mathscr{E}_0 + L}{(t_0+1)^2}, \end{aligned}$$

where c_p is a constant depending on p. Substituting the bounds of $\|F_v(t)\|$ and $\|F_x(t)\|$ into (42) to obtain that, for $n = 1, \dots, p$,

$$\left\| \nabla^n F(F_1, \dots, F_n) \right\| \leq c_{n,p} \frac{(1 + L + \mathscr{E}_0)^n}{t_0 + 1}, \tag{44}$$

where $c_{n,p}$ depends on n and p. Now we proceed to derive an upper bound for higher order time derivatives. By Lemma 10 we can write $\| \frac{\partial^n \varphi_h(y_0)}{\partial h^n} \| = \|F^{(n-1)}(\varphi_h(y_0))\| = \| \sum_{|\tau|=n} \alpha(\tau) F(\tau)(\varphi_h(y_0)) \|$. Hence, there exists a constant C depending on n and p, such that

$$\left\| \frac{d^n \varphi_h(y_0)}{dh^n} \right\| \leq \frac{C_0(1 + L + \mathscr{E}_0)^n}{t_0 + 1} \leq C_0(1 + L + \mathscr{E}_0)^n.$$

Similarly, by Lemma 11, we have the following equation

$$\frac{d^n \Phi_h(y_0)}{dh^n} = \sum_{i=1}^{s} b_i \left[h \frac{d^n F(z_i)}{dh^n} + n \frac{d^{n-1} F(z_i)}{dh^{n-1}} \right].$$

Hence, $\frac{d^{(n)} F(z_i)}{dh^n}$ has the same recursive tree structure as $F^{(n)}(y)$, except that we need to replace all F in the expression by $\frac{dz_i}{dh}$ and all $\nabla^n F(y)$ by $\nabla^n F(z_i)$. By Definition 1, we know that

$$\left\| \frac{dz_{i1}}{dh} \right\| \leq \sum_{j=1}^{S} |a_{ij}| \cdot \frac{\hat{c}_1(1 + \mathscr{E}_0 + L)}{(t_0+1)^2},$$

$$\left\| \frac{dz_{i2}}{dh} \right\| \leq \sum_{j=1}^{S} |a_{ij}| \cdot \frac{p+1}{t_0+1} \hat{c}_2(\mathscr{E}_0 + 1), \ \left\| \frac{dz_{i3}}{dh} \right\| \leq \Big| \sum_{j=1}^{S} a_{ij} \Big|.$$

Because $\|\nabla^n F(z_i)\|$ has the same boundary as $\|\nabla^n F(y)\|$, we also can get a boundary of $\|\frac{\partial^n F(z_i)}{\partial h^n}\|$ by the same argument with the boundedness of $\|\frac{\partial^n \varphi_h(y_0)}{\partial h^n}\|$, which is a constant determined by the integrator. We conclude that

$$\begin{aligned}\frac{\partial^n \Phi_h(y_0)}{\partial h^n} &= \frac{C_1(1+L+\mathscr{E}_0)^n + C_1' h(1+L+\mathscr{E}_0)^{n-1}}{t_0+1} \\ &\leq C_1(1+L+\mathscr{E}_0)^n + C_1' h(1+L+\mathscr{E}_0)^{n-1},\end{aligned}$$

where the constants C_1 and C_1' are determined by n, p and the integrator. ■

Lemma 23: For any $n \geq 1$, $\|\nabla^n \mathscr{E}(y)\| \leq C_p(t_0+1)^p(1+L+\mathscr{E}_0)$, where C_p is a constant depending on p.

Proof: By explicitly computing its entries $\frac{\partial^n \mathscr{E}}{\partial v^n}$, $\frac{\partial^n \mathscr{E}}{\partial x^n}$ and $\frac{\partial^n \mathscr{E}}{\partial t^n}$, we bound $\nabla^n \mathscr{E}(y)$ by

$$\begin{aligned}\|\nabla^n \mathscr{E}\| \leq \left\|\frac{\partial^n \mathscr{E}}{\partial v^n}\right\| + \left\|\frac{\partial^n \mathscr{E}}{\partial x^n}\right\| + \left\|\frac{\partial^n \mathscr{E}}{\partial t^n}\right\| + \sum_{l=1}^{p}\left(\left\|\frac{\partial^n \mathscr{E}}{\partial t^l \partial x^{n-l}}\right\|\right. \\ \left. + \left\|\frac{\partial^n \mathscr{E}}{\partial t^l \partial v^{n-l}}\right\| + \left\|\frac{\partial^n \mathscr{E}}{\partial v^l \partial x^{n-l}}\right\|\right). \quad (45)\end{aligned}$$

When $n > p$, from the definition of $\mathscr{E}$, we have

$$\begin{aligned}\|\nabla^n \mathscr{E}\| &\leq \sum_{l=0}^{p}\left\|\frac{\partial^n \mathscr{E}}{\partial t^l \partial x^{n-l}}\right\| \\ &\leq \sum_{l=0}^{p}\frac{p!}{(p-l)!}(t+1)^{p-l}\left\|\nabla^{n-l} f(x)\right\| \\ &\leq \sum_{l=0}^{p}\frac{p!}{(p-l)!}(t+1)^{p-l} L \\ &\leq (p+1)! L (t+1)^p,\end{aligned}$$

where the third inequality uses $\|\nabla^i f(x)\| \leq L$. Because $t+1 \leq c_1(t_0+1)$, we have that

$$\|\nabla^n \mathscr{E}\| \leq (p+1)! L c_1^p (t_0+1)^p \leq \hat{C}_p(t_0+1)^p(L+\mathscr{E}_0+1), \quad (46)$$

where $\hat{C}_p = (p+1)! c_1^p$ depends on p.

Similarly, for $n \leq p$, there is a similar bound, that is, $\|\nabla^n \mathscr{E}\| \leq \bar{C}_p(t_0+1)^p(L+\mathscr{E}_0+1)$ where the constant $\bar{C}_p$ depends on p and n. Let $C_p = \max\{\hat{C}_p, \bar{C}_p\}$ and then the proof of the lemma is completed. ■

Lemma 24: For any $n > 1$,

$$\left\|\frac{d^n \mathscr{E}(\varphi_h(y_0))}{dh^n}\right\| \leq C_{p,n}(1+L+\mathscr{E}_0)^{n+1},$$

where the constant $C_{p,n}$ depends on n and p.

Proof: By the chain rule, we have

$$\begin{aligned}&\frac{d^n \mathscr{E}(\varphi_h(y_0))}{dh^n} \\ &\quad = \sum_{k_1,\ldots,k_n}\frac{n!}{k_1!\cdots k_n!}\nabla^k \mathscr{E}(y)\, \Pi_{i=1}^n\left(\frac{d^i(\varphi_h(y_0))}{dh^i}\frac{1}{i!}\right)^{k_i},\end{aligned}$$

where the sum takes over $\{k_1,\ldots,k_n \in \mathbb{Z}_{\geq 0} | \sum_{i=1}^n ik_i = n\}$ and $k = \sum_{i=1}^n k_i$.
Then from (40), we have

$$\begin{aligned}&\Pi_{i=1}^n\left(\frac{d^i(\varphi_h(y_0))}{dh^i}\frac{1}{i!}\right)^{k_i} \\ &\quad\leq \Pi_{i=1}^n\left(\frac{C_0(1+L+\mathscr{E}_0)^i}{t_0+1}\frac{1}{i!}\right)^{k_i} \\ &\quad= \frac{1}{(1!)^{k_1}(2!)^{k_2}\cdots(n!)^{k_n}}\frac{(1+L+\mathscr{E}_0)^{k_1+2k_2+\cdots+nk_n}}{(t_0+1)^{k_1+k_2+\cdots+k_n}} \\ &\quad= \frac{1}{(1!)^{k_1}(2!)^{k_2}\cdots(n!)^{k_n}}\frac{(1+L+\mathscr{E}_0)^n}{(t_0+1)^k}.\end{aligned}$$

From Lemma 23, $\|\nabla^n \mathscr{E}\| \leq C_p(t_0+1)^p(L+\mathscr{E}_0+1)$ and then we have

$$\begin{aligned}&\frac{d^n \mathscr{E}(\varphi_h(y_0))}{dh^n} \\ &\quad= C_p(t_0+1)^p(1+L+\mathscr{E}_0)^{n+1} \\ &\qquad\cdot\sum_{k_1,\ldots,k_n}\frac{n!}{k_1!\cdots k_n!}\frac{1}{(1!)^{k_1}(2!)^{k_2}\cdots(n!)^{k_n}}\frac{1}{(t_0+1)^k} \\ &\quad\leq C_{p,n}(1+L+\mathscr{E}_0)^{n+1},\end{aligned}$$

where $C_{p,n} = C_p n^2 (t_0+1)^p$ is a positive constant. ■

Lemma 25: For any $n > 1$, $\left\|\frac{d^n \mathscr{E}(\Phi_h(y_0))}{dh^n}\right\| \leq C_{p,n}'(1+L+\mathscr{E}_0)^{n+1}$, where $C_{p,n}'$ is similar to $C_{p,n}$ in Lemma 24.

Proof: The proof is similar to Lemma 24. The difference is that instead of using (40), we use (41) to bound the high order derivatives of the trajectories. ■

V. NUMERICAL EXPERIMENTS

In this section, we perform a series of numerical experiments to verify the performance of the proposed scheme for minimizing strongly convex functions. First, when $p = 2$, we consider the effectiveness of ImRK with different objective functions, and compare it with GD as well as NAG. Then, with the same objectives, we considered the ODE (6) for different $p \geq 2$. Each experiment was repeated 10 times with different W, H and initial value conditions and averaged. For each method tested, we empirically choose the step size among $\{10^{-k} \, | k \in Z\}$. All figures in this section are on log-log scale.

A. OBJECTIVE FUNCTIONS

We first verify the convergence of the algorithm for fixed p and different integrators. The theoretical results show that for the gradual increase of s, the corresponding ImRK has an increase of q, which finally realizes the accelerated convergence of the algorithm. By minimizing a quadratic convex function of the form $f_1(x) = \|Wx - H\|^2$, the ODE (5) with $p = 2$ is simulated, i.e.,

$$\ddot{x}(t) + \frac{5}{t+1}\dot{x}(t) + 4\nabla f(x(t)) = 0,$$

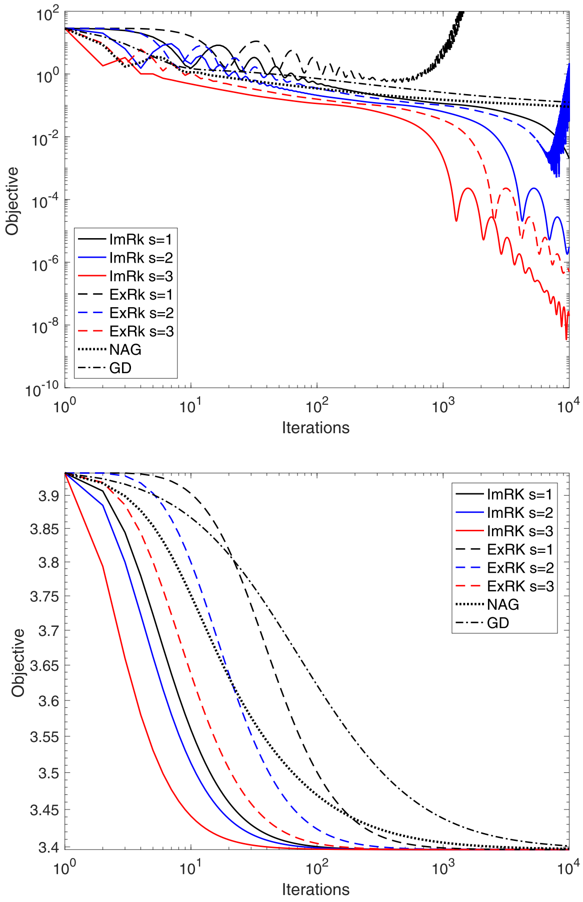

FIGURE 1. **Convergence paths of GD, NAG, ExRK and the proposed method with integrators of stages $s = 1$, $s = 2$ and $s = 3$. Top: The objective is a quadratic function f_1. Bottom: The objective is a logistic regression f_2.**

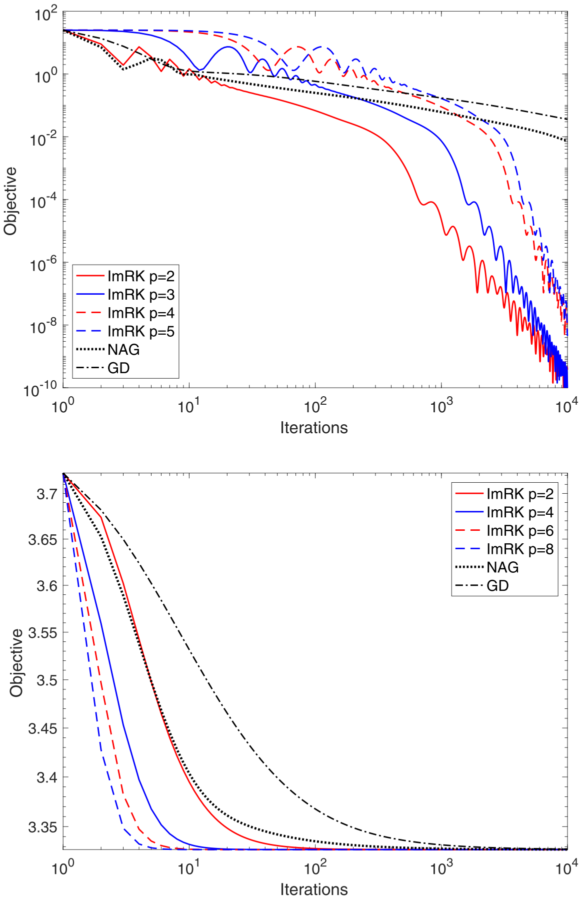

FIGURE 2. **Experiment results for the cases that Assumption 13 holds for $p \geq 2$. Minimizing the objective by simulating different ODEs with the 2-stage 4-order ImRK integrator. Top: the objective is a quadratic function. Bottom: the objective is a logistic regression.**

where $W \in \mathbb{R}^{10\times 10}$ and $H \in \mathbb{R}^{10}$. The entries of H are randomly selected from value 0 or 1. Each row W_i of W are generated by an i.i.d multivariate Gaussian distribution. Note that the quadratic objective $f_1(x)$ satisfies Assumption 13 with $p = 2$. The convergence paths of GD, NAG, ExRK and the proposed ImRK discretization procedure for minimizing the quadratic function $f_1(x)$ are demonstrated in the top subfigure of Figure 1. For the proposed method we consider the A-stable integrators with different stages, i.e., $s \in \{1, 2, 3\}$. We observe that GD reaches a exponential convergence rate, which verifies the theoretical convergence rate of GD when the function is strongly convex. NAG displays local acceleration when setting the optimal point as mentioned in [19]. Our theoretical results suggest that the convergence rate for $s \in \{1, 2, 3\}$ is $O((1 - \tilde{C}_{p,q}\frac{\mu}{L})^N N^{-2})$ for some constant C, which is faster than $O(N^{-2})$. At the same time, as shown in the top subfigure of Figure 1, the ImRK discretization algorithm actually achieves a convergence rate faster than $O(N^{-2})$. As a second example, we consider the logistic regression function $f_2(x) = \sum_{i=1}^{10} \log\left(1 + e^{-H_i x^T W_i}\right)$ which is convex and Lipschitz smooth. The used data points are generated in the same way as before. As shown in Section II, it satisfies Assumption 13 for arbitrary $p \geq 2$. The convergence results for f_2 are demonstrated in the bottom subfigure of Figure 1. It is observed that the ImRK tends to converge faster than the explicit method, which is consistent with our theoretical result in this paper.

B. DIFFERENT ODES

In this section, we discretize the ODE (6) with the objective functions that satisfy Assumption 13 with $p \geq 2$. However, it should be noted that for the quadratic objective function f_1, it is easy to known from (6) that when the parameter p changes, the dynamic system may become a stiff system. In general, the discrete-time algorithms, such as the explicit Runge-Kutta methods, are invalid for the stiff problem. In this paper, the A-stable implicit Runge-Kutta methods can effectively solve the above defects.

For all ODE discretization algorithms, we use an A-stable 2-stage 4-order ImRK integrator that calls the gradient oracle twice per iteration. In particular, we use the above quadratic objective $f_1(x) = \|Wx - H\|^2$ and $f_2(x) = \sum_{i=1}^{10} \log\left(1 + e^{-H_i x^T W_i}\right)$. We run all algorithms for the quadratic function and the logistic regression function with 10^4 iterations. We simulate the ODE (6) with different p values using the same numerical integrator and the same step size. Figure 2 summarizes the experimental results. For the quadratic objective function, we observe that when $p = 2$, the convergence of ImRK methods is faster than NAG. And it is interesting that when $p = 3$, the discretization is still stable with the convergence rate $O((1-\tilde{C}_{p,q}\frac{\mu}{L})^N N^{-3})$. When $p = 4$ and $p = 5$, although the discretization is stable, the convergence rate is lower than $O(N^{-4})$ and $O(N^{-5})$, respectively. The reason for this is that the term $(1 - \tilde{C}_{p,q}\frac{\mu}{L})^N$ on the right side of (28) plays a major role in the convergence of the algorithm, especially, that $\tilde{C}_{p,q}$ decreases with the increase of p makes the convergence rate slow. On the contrary, for the logistic regression function f_2, this kind of phenomenon disappears. On the one hand, it is because the ODE corresponding to the function f_2 has good properties. On the other hand, it is because the use of the A-stable method restrain the divergence effectively. By Theorem 21, if we set $p = 2$, we can achieve a faster convergence rate than $O(N^{-2})$. We run the experiments with different p values and summarize the results in the bottom subfigure of Figure 2. Note that when $q > 2$, the convergence of ImRK methods is faster than NAG.

VI. CONCLUSION AND DISCUSSION

In this paper, we have demonstrated the effective integration of implicit Runge-Kutta method and Bregman Lagrange equation in obtaining accelerated optimization methods. By applying the A-stable implicit Runge-Kutta method to the ODE derived by the Bregman Lagrangian, we found that our discrete-time algorithm are provably faster than other algorithms, such as gradient and Nesterov's accelerated gradient descent and the convergence rate is closely related with the order of integrator is higher as well as the larger range of step size h.

The future work is to extend the existing results to more general hybrid algorithms, such as explicit-implicit Runge-Kutta method, multi-stage high-order implicit Runge-Kutta method and the general form of multi-level multi-stage method.

ACKNOWLEDGMENT

The authors would like to thank Dr. Y. Yuan for interesting and useful discussions on the convex optimization methods.

REFERENCES

[1] C. M. Bishop, *Pattern Recognition and Machine Learning*. New York, NY, USA: Springer, 2006.

[2] L. Ljung, *System Identification: Theory for the User*. Upper Saddle River, NJ, USA: Prentice-Hall, 1987.

[3] K. Zhou, J. C. Doyle, and K. Glover, *Robust and Optimal Control*, vol. 40. Upper Saddle River, NJ, USA: Prentice-Hall, 1996.

[4] A. E. Bryson, *Applied Optimal Control: Optimization, Estimation and Control*. Evanston, IL, USA: Routledge, 2018.

[5] D. Leonard, N. Van Long, and V. L. Ngo, *Optimal Control Theory and Static Optimization in Economics*. Cambridge, U.K.: Cambridge Univ. Press, 1992.

[6] S. Boyd and L. Vandenberghe, *Convex Optimization*. Cambridge, U.K.: Cambridge Univ. Press, 2004.

[7] B. Polyak, "Some methods of speeding up the convergence of iteration methods," *USSR Comput. Math. Math. Phys.*, vol. 4, no. 5, pp. 1–17, Jan. 1964.

[8] B. Polyak, *Introduction to Optimization*, vol. 1. New York, NY, USA: Optimization Software, Publications Division, 1987.

[9] Y. Nesterov, "A method of solving a convex programming problem with convergence rate $\mathcal{O}(1/k^2)$," *Soviet Math. Doklady*, vol. 27, no. 2, pp. 372–376, 1983.

[10] Y. Nesterov, *Introductory Lectures on Convex Optimization: A Basic Course*, vol. 87. Springer, 2013.

[11] A. Beck and M. Teboulle, "A fast iterative shrinkage-thresholding algorithm for linear inverse problems," *SIAM J. Imag. Sci.*, vol. 2, no. 1, pp. 183–202, Jan. 2009.

[12] Y. Nesterov, "Gradient methods for minimizing composite functions," *Math. Program.*, vol. 140, no. 1, pp. 125–161, Aug. 2013.

[13] Y. Nesterov, "Efficiency of coordinate descent methods on huge-scale optimization problems," *SIAM J. Optim.*, vol. 22, no. 2, pp. 341–362, Jan. 2012.

[14] Q. Lin, Z. Lu, and L. Xiao, "An accelerated proximal coordinate gradient method," in *Proc. Adv. Neural Inf. Process. Syst.*, 2014, pp. 3059–3067.

[15] C. Hu, W. Pan, and J. T. Kwok, "Accelerated gradient methods for stochastic optimization and online learning," in *Proc. Adv. Neural Inf. Process. Syst.*, 2009, pp. 781–789.

[16] G. Lan, "An optimal method for stochastic composite optimization," *Math. Program.*, vol. 133, nos. 1–2, pp. 365–397, Jun. 2012.

[17] I. Necoara, Y. Nesterov, and F. Glineur, "Linear convergence of first order methods for non-strongly convex optimization," *Math. Program.*, vol. 175, nos. 1–2, pp. 69–107, May 2019.

[18] A. Juditsky and A. Nemirovski, "First order methods for nonsmooth convex large-scale optimization, I: General purpose methods," in *Optimization for Machine Learning*. Cambridge, MA, USA: MIT Press, 2011, pp. 121–148.

[19] W. Su, S. Boyd, and E. Candes, "A differential equation for modeling Nesterov's accelerated gradient method: Theory and insights," in *Proc. Adv. Neural Inf. Process. Syst.*, 2014, pp. 2510–2518.

[20] B. Shi, S. S. Du, M. I. Jordan, and W. J. Su, "Understanding the acceleration phenomenon via high-resolution differential equations," 2018, *arXiv:1810.08907*. [Online]. Available: https://arxiv.org/abs/1810.08907

[21] B. Shi, S. S. Du, W. J. Su, and M. I. Jordan, "Acceleration via symplectic discretization of high-resolution differential equations," 2019, *arXiv:1902.03694*. [Online]. Available: https://arxiv.org/abs/1902.03694

[22] A. Wibisono, A. C. Wilson, and M. I. Jordan, "A variational perspective on accelerated methods in optimization," *Proc. Nat. Acad. Sci. USA*, vol. 113, no. 47, pp. E7351–E7358, Nov. 2016.

[23] Z. Allen-Zhu and L. Orecchia, "Linear coupling: An ultimate unification of gradient and mirror descent," 2014, *arXiv:1407.1537*. [Online]. Available: https://arxiv.org/abs/1407.1537

[24] J. Zhang, S. Sra, and A. Jadbabaie, "Acceleration in first order quasi-strongly convex optimization by ODE discretization," 2019, *arXiv:1905.12436*. [Online]. Available: https://arxiv.org/abs/1905.12436

[25] J. Zhang, A. Mokhtari, S. Sra, and A. Jadbabaie, "Direct Runge–Kutta discretization achieves acceleration," in *Proc. Adv. Neural Inf. Process. Syst.*, 2018, pp. 3904–3913.

[26] M. Muehlebach and M. I. Jordan, "A dynamical systems perspective on Nesterov acceleration," 2019, *arXiv:1905.07436*. [Online]. Available: https://arxiv.org/abs/1905.07436

[27] Y. Yuan, M. Li, J. Liu, and C. Tomlin, "On the powerball method: Variants of descent methods for accelerated optimization," *IEEE Control Syst. Lett.*, vol. 3, no. 3, pp. 601–606, Jul. 2019.

[28] E. Hairer, S. P. Nørsett, and G. Wanner, *Solving Ordinary Differential Equations I: Nonstiff Problems*. Berlin, Germany: Springer, 1987.

[29] G. Wanner and E. Hairer, *Solving Ordinary Differential Equations II: Stiff and Differential-Algebraic Problems*. Berlin, Germany: Springer, 1996.

[30] S. F. Li, *Theory of Computational Methods for Stiff Differential Equations*. Changsha, China: Hunan Science and Technology Publisher, 1997.

[31] G. G. Dahlquist, "A special stability problem for linear multistep methods," *BIT Numer. Math.*, vol. 3, no. 1, pp. 27–43, 1963.

[32] E. Hairer, C. Lubich, and G. Wanner, *Geometric Numerical Integration: Structure-Preserving Algorithms for Ordinary Differential Equations*, vol. 31. Berlin, Germany: Springer, 2006.

RUIJUAN CHEN received the B.S. degree in applied mathematics from Xuchang University, Xuchang, China, in 2014, and the M.S. degree in computational mathematics from the Huazhong University of Science and Technology, Wuhan, China, in 2016, where she is currently pursuing the Ph.D. degree with the School of Artificial Intelligence and Automation. Her research interest includes the design and the theoretical analysis of stochastic optimization algorithms.

XIUTING LI received the B. Eng. degree in mathematics and statistics from Hubei Engineering University, Xiaogan, China, in 2011, and the Ph.D. degree in mathematics and statistics from the Huazhong University of Science and Technology, Wuhan, China, in 2016. She is currently a Postdoctoral Researcher with the School of Artificial Intelligence and Automation, Huazhong University of Science and Technology. She has published more than ten articles in refereed international journals. She has been involved in some projects granted by the National Science Foundation Committee of China in recent years. Her current research interests include modeling, analysis, and the controller design of distributed parameter systems.