Continuous-Time Analysis of Accelerated Gradient Methods via Conservation Laws in Dilated Coordinate Systems

Jaewook J. Suh [1] **Gyumin Roh** [1] **Ernest K. Ryu** [1]

Abstract

We analyze continuous-time models of accelerated gradient methods through deriving conservation laws in dilated coordinate systems. Namely, instead of analyzing the dynamics of $X(t)$, we analyze the dynamics of $W(t) = t^\alpha(X(t) - X_c)$ for some α and X_c and derive a conserved quantity, analogous to physical energy, in this dilated coordinate system. Through this methodology, we recover many known continuous-time analyses in a streamlined manner and obtain novel continuous-time analyses for OGM-G, an acceleration mechanism for efficiently reducing gradient magnitude that is distinct from that of Nesterov. Finally, we show that a semi-second-order symplectic Euler discretization in the dilated coordinate system leads to an $\mathcal{O}(1/k^2)$ rate on the standard setup of smooth convex minimization, without any further assumptions such as infinite differentiability.

1. Introduction

Despite the significance of acceleration within the study of first-order optimization methods, a fundamental understanding of the acceleration phenomena remains elusive. Recently, continuous-time analyses of accelerated gradient methods have been extensively pursued, even using ideas from mathematical physics. However, these continuous-time analyses still retain a component of mystery: They rely on establishing that certain energy functions are nonincreasing but do not justify the origin of such energy functions.

In this work, we present a methodology for analyzing accelerated gradient methods through deriving a conservation law, analogous to the conservation of energy of physics, in a dilated coordinate system. Namely, instead of analyzing the dynamics of $X(t)$, we analyze the dynamics of $W(t) = t^\alpha(X(t) - X_c)$ for some $\alpha \in \mathbb{R}$ and $X_c \in \mathbb{R}^n$.

Through this methodology, we recover many known continuous-time analyses in a streamlined manner. Furthermore, the methodology enables us to perform a novel analysis of an ODE model of OGM-G of Kim & Fessler (2021), an acceleration mechanism distinct from that of (Nesterov, 1983). Finally, we show that a semi-second-order symplectic Euler discretization in the dilated coordinate system leads to an $\mathcal{O}(1/k^2)$ rate on the standard setup of smooth convex minimization, without any further assumptions such as infinite differentiability.

1.1. Preliminaries and notation

We review the standard definitions of convex optimization and set up the notation (Nesterov, 2004; Boyd & Vandenberghe, 2004; Bauschke & Combettes, 2017; Nesterov, 2018; Ryu & Yin, 2022). Throughout the paper, we use $\mathbb{R}^n$ for the underlying Euclidean space with Euclidean norm $\|\cdot\|$ and inner product $\langle\cdot,\cdot\rangle$. For $L > 0$, $f\colon \mathbb{R}^n \to \mathbb{R}$ is L-smooth if f is differentiable and

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

For $\mu > 0$, $g\colon \mathbb{R}^n \to \mathbb{R}$ is μ-strongly convex if $g(x) - (\mu/2)\|x\|^2$ is convex. When f is differentiable and convex,

$$f(x) - f(y) - \langle \nabla f(y), x - y\rangle \geq 0$$

holds for all $x, y \in \mathbb{R}^n$, and we refer to this inequality as the *convexity inequality*. Throughout this paper, consider

$$\underset{x\in\mathbb{R}^n}{\text{minimize}} \quad f(x), \tag{1}$$

where $f\colon \mathbb{R}^n \to \mathbb{R}$ is convex and differentiable. When (1) has a minimizer, write $X_\star$ to denote a minimizer. Write $f_\star = \inf_{x\in\mathbb{R}^n} f(x)$ for the optimal value of the problem.

Energy and conservation law. Let $A\colon (0,\infty) \to \mathbb{R}$ be differentiable and $B\colon (0,\infty) \to \mathbb{R}$ be integrable. Suppose

$$0 = \dot{A}(t) + B(t)$$

[1]Department of Mathematical Sciences, Seoul National University, Seoul, Korea. Correspondence to: Ernest K. Ryu <ernestryu@snu.ac.kr>.

Proceedings of the 39th International Conference on Machine Learning, Baltimore, Maryland, USA, PMLR 162, 2022.

holds for all $t > 0$. Then, for $0 < t_0 < t < \infty$, integrating from t_0 to t gives us the *conservation law*

$$E \equiv A(t_0) = A(t) + \int_{t_0}^{t} B(s)\, ds,$$

where the *energy* E is independent of time. Moreover, if the limit $\lim_{t_0 \to 0} A(t_0)$ exists, then

$$E \equiv \lim_{t_0 \to 0} A(t_0) = A(t) + \int_{0}^{t} B(s)\, ds.$$

Partial derivatives. Consider a function $U(W, t)$ with variables $W = (w_1, \ldots, w_n) \in \mathbb{R}^n$ and $t \in \mathbb{R}$. Define

$$\nabla_W U(W,t) = \left(\frac{\partial}{\partial w_1} U(W,t), \ldots, \frac{\partial}{\partial w_n} U(W,t) \right) \in \mathbb{R}^n.$$

When $W(t)$ is differentiable, the chain rule gives us

$$\frac{d}{dt} U(W(t),t) = \left\langle \nabla_W U(W(t),t), \dot{W}(t) \right\rangle + \frac{\partial}{\partial t} U(W(t),t). \tag{2}$$

To clarify, the distinction between $\frac{d}{dt}$ and $\frac{\partial}{\partial t}$ corresponds to viewing $W(t)$ as a curve dependent on t or viewing W as an input to U independent of t. We clarify this notation fully in Appendix A. Then for $0 < t_0 < t < \infty$, integrating from t_0 to t gives us

$$\int_{t_0}^{t} \left\langle \nabla_W U(W,s), \dot{W}(s) \right\rangle \, ds$$
$$= U(W(t),t) - U(W(t_0),t_0) - \int_{t_0}^{t} \frac{\partial}{\partial s} U(W,s)\, ds.$$

1.2. Prior work

In convex optimization and machine learning, the classical goal is to reduce the function value efficiently. In the smooth convex setup, Nesterov's celebrated accelerated gradient method (AGM) (Nesterov, 1983) achieves an accelerated rate of $\mathcal{O}(1/k^2)$. Recently, the optimized gradient method (OGM) (Kim & Fessler, 2016) improved the rate of AGM by a factor of 2, and this rate is in fact exactly optimal (Drori, 2017). In the smooth strongly convex setup, the strongly convex AGM (SC-AGM) (Nesterov, 2018, 2.2.22) achieves an accelerated rate. The review by d'Aspremont et al. (2021) provides a comprehensive historical review.

The study of first-order convex optimization algorithms efficiently reducing the squared gradient norm was initiated by Nesterov (2012). For smooth non-convex minimization, gradient descent (GD) achieves an $\mathcal{O}((f(x_0) - f_\star)/k)$ rate (Nemirovski, 1999, Proposition 3.3.1). In the smooth convex setup, OGM-G (Kim & Fessler, 2021) achieves an $\mathcal{O}(f(x_0) - f_\star)/k^2)$ rate. M-OGM-G (Zhou et al., 2022) and OBL-G$_\flat$ (Park & Ryu, 2021) are variants of OGM-G achieving similar rates. Combining AGM with OGM-G (Nesterov, 2018, Remark 2.1) yields an $\mathcal{O}(\|x_0 - x_\star\|^2 / k^4)$ rate, which matches the $\Omega(\|x_0 - x_\star\|^2 / k^4)$ lower bound of (Nemirovsky, 1991; 1992) and is therefore optimal.

An ODE model for the heavy ball method with constant friction, i.e., constant damping, was introduced by Polyak (1964) and follow-up work studying variations flourished (Attouch & Alvarez, 1998; Alvarez & Attouch, 2001; Attouch & Czarnecki, 2002; Alvarez et al., 2002; Attouch et al., 2002; 2012; Attouch & Czarnecki, 2017; Boţ & Csetnek, 2017; 2019; Adly & Attouch, 2020b; Adly et al., 2021b; Aujol et al., 2021; 2022). The study of ODE models of AGM and accelerated mirror descent with vanishing damping was initiated by Su et al. (2014; 2016); Krichene et al. (2015). Specifically, Su et al. (2014) studied the dynamics of $0 = \ddot{X} + \frac{r}{t}\dot{X} + \nabla f(X)$ and proved $f(X(t)) - f_\star \leq (r-1)^2 \|X_0 - X_\star\|^2 / (2t^2)$ for $r \geq 3$. Attouch et al. (2018c) improved the constant of this bound for $r > 3$. For $r < 3$, Attouch et al. (2019c) established an $\mathcal{O}(t^{-2r/3})$ rate. Improved rates under the additional, so-called, $\mathbf{H}_1(\gamma)$ hypothesis were established by Aujol et al. (2019); Sebbouh et al. (2019); Apidopoulos et al. (2021). A wide range of variations of the ODE with vanishing damping were also studied (Attouch & Chbani, 2015; May, 2017; Attouch et al., 2018b;d; Attouch & Cabot, 2018a; Attouch et al., 2019b; Attouch & Peypouquet, 2019; Attouch & László, 2020; Attouch et al., 2020a; 2021a;d; Attouch & László, 2021; Attouch & Cabot, 2017; Attouch & Laszlo, 2021; Boţ et al., 2021; Attouch et al., 2022; 2021b). Similar analyses were extended to differential inclusions for non-differentiable functions (Attouch & Maingé, 2011; Attouch & Peypouquet, 2016; Aujol & Dossal, 2017b; Apidopoulos et al., 2017; 2018), monotone inclusions (Boţ & Csetnek, 2016; 2018; Boţ et al., 2018; Bot & Hulett, 2022), primal-dual methods (Boţ & Nguyen, 2021), and splitting methods França et al. (2018); Hassan-Moghaddam & Jovanović (2021); França et al. (2021b); Attouch et al. (2021c).

This intense study of ODEs modeling optimization algorithms motivated the development of tools utilizing the following ideas: variational principle and Lagrangian mechanics (Wibisono et al., 2016; Jordan, 2018; Zhang et al., 2021; Wilson et al., 2021); duality gap and convex-analytical techniques (Diakonikolas & Orecchia, 2019); Hamiltonian mechanics (Diakonikolas & Jordan, 2021); control theory (Hu & Lessard, 2017); continuous-time complexity lower bounds (Muehlebach & Jordan, 2020); and perturbation analysis of physics, leading to the high-resolution ODE (Shi et al., 2021).

The study of continuous-time models, in turn, motivated the study of discretizing such ODEs to obtain implementable algorithms. Discretizing ODEs with vanishing damping

(Wibisono et al., 2016; Attouch et al., 2018a; Attouch & Cabot, 2018b; Attouch et al., 2019a; 2020a; Adly & Attouch, 2020a; Attouch & Cabot, 2020; Attouch et al., 2020b; Adly & Attouch, 2021; Adly et al., 2021a;c; Diakonikolas & Jordan, 2021) and discretizing alternate ODEs (Scieur et al., 2017; Wilson et al., 2019; Muehlebach & Jordan, 2019; Zhang et al., 2019) have been studied. Specifically, Zhang et al. (2018) achieved an $\mathcal{O}(1/k^2)$ rate using the Runge–Kutta discretization on the ODE by Su et al. (2014) under additional assumptions.

The study of using symplectic integrators, a discretization scheme designed to conserve energy (Hairer et al., 2006), for discretizing the ODE models was initiated by Betancourt et al. (2018) and was further developed in a series of work (Maddison et al., 2018; França et al., 2020a;b; Muehlebach & Jordan, 2021; França et al., 2021a). However, these approaches did not obtain an asymptotic $\mathcal{O}(1/k^2)$ rate in the sense usually considered in optimization. An $\mathcal{O}(1/k^2)$ rate was obtained by Shi et al. (2019) combining symplectic integration with the high-resolution ODE framework.

Recently, Even et al. (2021) introduced the "continuized" framework of accelerated gradient methods, which uses a stochastic jump process to perform randomized discretizations. The framework can utilize the simpler continuous-time analysis while producing an implementable (but randomized) discrete algorithm with rate $\mathcal{O}(1/k^2)$.

1.3. Contribution

The central thesis, the main contribution, of this paper is that continuous-time analyses of accelerated gradient methods significantly simplify under an alternate dilated coordinate system. We establish this claim by presenting a methodology analyzing the ODEs by deriving conservation laws in dilated coordinate systems and recovering many prior analyses in a streamlined manner. We then use the methodology to perform the first continuous-time analysis of OGM-G, whose acceleration mechanism was understood far less than the acceleration mechanism of Nesterov.

Furthermore, we show that the coordinate change can also benefit the analysis of discretizations. Specifically, we apply a semi-second-order symplectic Euler discretization in the dilated coordinate system to obtain an $\mathcal{O}(1/k^2)$ rate in the standard setup of smooth convex minimization, without any further assumptions such as infinite differentiability. This is the first result of its kind, in the precise sense clarified in Section 5.1, and it will be interesting to see, in future work, to what extent discretizations exploiting our dilated coordinates can achieve competitive rates.

2. Conservation laws from dilated coordinates

Our main methodology for continuous-time analysis is to perform a coordinate change and then obtain a conservation law. In this section, we quickly exhibit this methodology applied to the classical AGM ODE and then present a generalized form which we will use in later sections.

Consider problem (1). Assume a minimizer of f exists and write $X_\star$ for a minimizer of f. (We do not assume the minimizer is unique.) Write $f_\star = f(X_\star)$. The AGM ODE presented by Su et al. (2014) is

$$0 = \ddot{X} + \frac{3}{t}\dot{X} + \nabla f(X) \tag{3}$$

with initial condition $X(0) = X_0$, $\dot{X}(0) = 0$. Here, $X\colon [0,\infty) \to \mathbb{R}^n$ is a function of the time t, but we often write X in place of $X(t)$ for the sake of notational brevity. Consider the dilated coordinate $W = t^\alpha(X - X_\star)$ with a yet undetermined $\alpha \in \mathbb{R}$. The ODE in the W coordinate is

$$0 = \frac{1}{t^\alpha}\ddot{W} + \frac{3-2\alpha}{t^{\alpha+1}}\dot{W} + \nabla_W U(W,t) \tag{4}$$

with

$$U(W,t) = \frac{\alpha(\alpha-2)}{2t^{\alpha+2}}\|W\|^2 + t^\alpha\left(f\left(X(W,t)\right) - f_\star\right) \tag{5}$$

and $X(W,t) = \frac{W}{t^\alpha} + X_\star$. Since U contains $t^\alpha(f(X) - f_\star)$, we choose $\alpha = 2$ in anticipation of the $\mathcal{O}(1/t^2)$ rate to get

$$0 = \frac{1}{t^2}\ddot{W} - \frac{1}{t^3}\dot{W} + \nabla_W U(W,t). \tag{6}$$

Taking the inner product between $\dot{W}$ and (6) and using (2), we get

$$\begin{aligned} 0 &= \frac{d}{dt}\left(\frac{1}{2t^2}\left\|\dot{W}\right\|^2\right) + \left\langle \nabla_W U(W,t), \dot{W}(t)\right\rangle \\ &= \frac{d}{dt}\left(\frac{1}{2t^2}\left\|\dot{W}\right\|^2 + U(W(t),t)\right) - \frac{\partial}{\partial t}U(W(t),t). \end{aligned}$$

The corresponding conservation law is

$$\begin{aligned} E &\equiv 2\|X_0 - X_\star\|^2 \\ &= \lim_{t_0\to 0}\left(\frac{1}{2t_0^2}\left\|\dot{W}(t_0)\right\|^2 + U(W(t_0),t_0)\right) \\ &= \frac{1}{2t^2}\left\|\dot{W}(t)\right\|^2 + U(W(t),t) - \int_0^t \frac{\partial}{\partial s}U(W(s),s)\,ds. \end{aligned}$$

From $\frac{\partial}{\partial t}X(W,t) = -\frac{2}{t^3}W = -\frac{2}{t}(X - X_\star)$, we get

$$\begin{aligned} -\frac{\partial}{\partial t}U(W,t) &= -\frac{\partial}{\partial t}t^2\left(f\left(X(W,t)\right) - f_\star\right) \\ &= 2t\big(f_\star - f\left(X\right) - \langle \nabla f(X), X_\star - X\rangle\big) \end{aligned}$$

and

$$\begin{aligned} E &\equiv 2\left\|X_0 - X_\star\right\|^2 \\ &= t^2\left(f(X) - f_\star\right) + \frac{1}{2}\left\|t\dot{X} + 2(X - X_\star)\right\|^2 \\ &\quad + \int_0^t 2s\big(f_\star - f\left(X\right) - \langle \nabla f(X), X_\star - X\rangle\big)\, ds \end{aligned} \tag{7}$$

for all $t \geq 0$. Since f is convex, the integrand is nonnegative, and we conclude

$$f(X) - f_\star \leq \frac{E}{t^2} = \frac{2\left\|X_0 - X_\star\right\|^2}{t^2}.$$

General form of conservation laws. We now generalize the previous analysis for later sections. Let $U : \mathbb{R}^n \times \mathbb{R} \to \mathbb{R}$, and consider the ODE

$$0 = a(t)\ddot{W} + b(t)\dot{W} + \nabla_W U(W, t).$$

Take the inner product with $\dot{W}$ and integrate to obtain the conservation law

$$\begin{aligned} E &\equiv \frac{a(t_0)}{2}\left\|\dot{W}(t_0)\right\|^2 + U(W(t_0), t_0) \\ &= \frac{a(t)}{2}\left\|\dot{W}(t)\right\|^2 + \int_{t_0}^t \left(b(s) - \frac{\dot{a}(s)}{2}\right)\left\|\dot{W}(s)\right\|^2 ds \\ &\quad + U(W(t), t) - \int_{t_0}^t \frac{\partial}{\partial s} U(W(s), s)\, ds. \end{aligned} \tag{8}$$

Note that if $a(t) = 1$ and $U(W, t) = U(W)$, then this convservation law is nothing but the familiar conservation of energy in physics; within E, the first term $(1/2)\|\dot{W}\|^2$ is kinetic energy, the second term $\int_{t_0}^t b\|\dot{W}\|^2 ds$ is energy dissipated way as heat due to friction, the third term $U(W)$ is potential energy, and the fourth term vanishes as the potential U is independent of time.

Throughout this paper, we consider dilated coordinates of the form $W = e^{\gamma(t)}(X - X_c)$ for some $X_c \in \mathbb{R}^n$. As a consequence, $U(W, t)$ will contain $e^{\gamma(t)}(f(X(W, t)) - f(X_c))$. The convexity inequality enters the integral of $\frac{\partial}{\partial s}U(W, s)$ through the identity

$$\begin{aligned} &-\frac{\partial}{\partial t} e^{\gamma(t)}\left(f\left(X(W, t)\right) - f(X_c)\right) \\ &= \dot{\gamma}(t) e^{\gamma(t)}\big(f(X_c) - f\left(X\right) - \langle \nabla f(X), X_c - X\rangle\big). \end{aligned}$$

Note, if $e^{\gamma(t)} = 1$ for all t, i.e. if there is no coordinate change, then $\dot{\gamma}(t) = 0$ and the convexity inequality does not enter the conservation law. In this sense, the coordinate change is essential for our analysis to utilize convexity.

Connection with Lyapunov analyses. Our analyses based on conservation laws are not fundamentally different from the Lyapunov analyses of the prior work. The first two terms of the conservation law for the AGM ODE

$$\Phi(t) = t^2\left(f(X) - f_\star\right) + \frac{1}{2}\left\|t\dot{X} + 2(X - X_\star)\right\|^2,$$

form the exact Lyapunov function of Su et al. (2014). Once $\Phi(t)$ is stated, it is relatively straightforward to verify $\dot{\Phi}(t) \leq 0$ through direct differentiation. The conservation laws of Section 3 also contain Lyapunov functions of prior work (Attouch et al., 2019c; Aujol & Dossal, 2017a; Aujol et al., 2019).

The analyses of prior work often start by stating a Lyapunov function of unclear origin and then proceed with the analysis. In truth, these Lyapunov functions are obtained through many hours of trial and error. A core motivation of our work is to provide a systematic methodology for obtaining such Lyapunov functions.

The closely related prior work of Diakonikolas & Jordan (2021) presents a methodology based on Hamiltonian mechanics. While they also provide a unified methodology for analyzing continuous-time models of accelerated gradient methods, there are some key differences that we further clarify in Appendix B. One key difference is that while we start from a given ODE and derive conservations laws, Diakonikolas & Jordan (2021) start from a Hamiltonian with "potential energy" and a "kinetic energy" terms and derive the ODE. From our framework, a $\|W\|^2$ term arises naturally as in (5) and as in the third term of (10), but $\|W\|^2$ does not arise from the approach of Diakonikolas & Jordan (2021). Our analyses of the generalized AGM, SC-AGM, and OGM-G ODEs crucially rely on using the $\|W\|^2$ term and therefore cannot be obtained by the methodology of Diakonikolas & Jordan (2021) as is.

3. Continuous-time analyses of Nesterov-type acceleration via conservation laws in dilated coordinate systems

Again, consider problem (1). Assume a minimizer of f exists and write $X_\star$ for a minimizer of f. Write $f_\star = f(X_\star)$. Su et al. (2016) presented the generalized ODE

$$0 = \ddot{X} + \frac{r}{t}\dot{X} + \nabla f(X) \tag{9}$$

and provided Lyapunov analyses for $r \geq 3$. We consider the dilated coordinate $W = t^\alpha(X - X_\star)$ and follow a similar line of reasoning as that of Section 2 to obtain the

conservation law

$$
\begin{aligned}
E \equiv\ & t^{\alpha}\left(f\left(X\right)-f_{\star}\right)+\frac{1}{2}t^{\alpha-2}\left\|t\dot{X}+\alpha(X-X_{\star})\right\|^{2}\\
&+\frac{\alpha(\alpha+1-r)}{2}t^{\alpha-2}\left\|X-X_{\star}\right\|^{2} \qquad (10)\\
&+\int_{t_0}^{t}\left(\frac{(2r-3\alpha)s^{\alpha-3}}{2}\left\|s\dot{X}+\alpha(X-X_{\star})\right\|^{2}\right.\\
&\qquad\left.+\frac{\alpha(\alpha+1-r)(\alpha+2)}{2}s^{\alpha-3}\left\|X-X_{\star}\right\|^{2}\right)ds\\
&+\int_{t_0}^{t}\alpha s^{\alpha-1}\left(f_{\star}-f(X)-\langle\nabla f(X),X_{\star}-X\rangle\right)\,ds.
\end{aligned}
$$

Note that when $r=3$, $\alpha=2$, and $t_0=0$, half of the terms vanish and the conservation law reduces to (7).

Throughout this section, we present the analysis results based on conservation laws while deferring the detailed derivations to Appendix C.

3.1. AGM ODE $r>3$

Let $r>3$. Plug $\alpha=2$ and $t_0=0$ into (10) and evaluate integrals as described in Appendix C.2 to get

$$
\begin{aligned}
E &\equiv (5-r)\left\|X_0-X_{\star}\right\|^{2}\\
&=-2(r-3)\left\|X_0-X_{\star}\right\|^{2}\\
&\quad+t^{2}\left(f(X)-f_{\star}\right)+\frac{1}{2}\left\|t\dot{X}+2(X-X_{\star})\right\|^{2}\\
&\quad+(r-3)\left\|X-X_{\star}\right\|^{2}+\int_{0}^{t}\frac{r-3}{s}\left\|s\dot{X}\right\|^{2}ds\\
&\quad+\int_{0}^{t}2s\left(f_{\star}-f(X)-\langle\nabla f(X),X_{\star}-X\rangle\right)ds.
\end{aligned}
$$

All terms depending on t are nonnegative when $r>3$. Thus $E+2(r-3)\left\|X_0-X_{\star}\right\|^{2}\geq t^{2}(f(X)-f_{\star})$ holds, and we conclude

$$
f(X)-f_{\star}\leq\frac{(r-1)\left\|X_0-X_{\star}\right\|^{2}}{t^{2}}.
$$

This rate improves upon the rate $f(X)-f_{\star}\leq\frac{(r-1)^{2}\left\|X_0-X_{\star}\right\|^{2}}{2t^{2}}$ by Su et al. (2014) and matches the rate of Attouch et al. (2018c). This conservation law also implies $E\geq(r-3)\left\|X-X_{\star}\right\|^{2}$, and boundedness of $\left\|X-X_{\star}\right\|$ can be used to establish convergence of $X(t)$ (Chambolle & Dossal, 2015; Attouch et al., 2018c).

3.2. AGM ODE $r<3$

Let $0\leq r<3$. Plug $\alpha=\frac{2r}{3}$ to (10) to get

$$
\begin{aligned}
E &= t^{\frac{2r}{3}}\left(f\left(X\right)-f_{\star}\right)+\frac{r(3-r)}{9}t^{\frac{2r}{3}-2}\left\|X-X_{\star}\right\|^{2}\\
&\quad+\frac{1}{2}t^{\frac{2r}{3}-2}\left\|t\dot{X}+\frac{2r}{3}(X-X_{\star})\right\|^{2}\\
&\quad+\int_{t_0}^{t}\frac{2}{27}r(3-r)(3+r)s^{\frac{2r}{3}-3}\left\|X-X_{\star}\right\|^{2}ds\\
&\quad+\int_{t_0}^{t}\frac{2r}{3}s^{\frac{2r}{3}-1}\left(f_{\star}-f(X)-\langle\nabla f(X),X_{\star}-X\rangle\right)\,ds.
\end{aligned}
$$

We let the starting time be nonzero, i.e., $t_0>0$, to ensure all of the terms do not blow up. All terms are nonnegative. Thus $E\geq t^{\frac{2r}{3}}\left(f(X)-f_{\star}\right)$, and we conclude

$$
f(X)-f_{\star}\leq\frac{E}{t^{\frac{2r}{3}}}.
$$

This recovers the result of Attouch et al. (2019c).

3.3. AGM ODE with growth condition

Aujol et al. (2019) consider convex functions satisfying the so-called "$\mathbf{H}_1(\gamma)$ hypothesis", defined as

$$
f(x)-f_{\star}\leq\frac{1}{\gamma}\left\langle\nabla f(x),x-X_{\star}\right\rangle,\qquad\forall x\in\mathbb{R}^{n}
$$

for a $\gamma\geq1$, and obtain improved rates. To utilize the $\mathbf{H}_1(\gamma)$ hypothesis, rather than the convexity inequality, we rescale the ODE by multiplying t^{β} and then obtain the conservation law (8) with the rescaled ODE. The derivations are detailed in Appendix C.3. With values $\alpha=\frac{2r}{\gamma+2}$ and $\beta=\frac{2(\gamma-1)r}{\gamma+2}$ we get

$$
\begin{aligned}
E\equiv\ & t^{\frac{2\gamma r}{\gamma+2}}(f(X)-f_{\star})+\frac{1}{2}t^{\frac{2\gamma r}{\gamma+2}-2}\left\|t\dot{X}+\alpha(X-X_{\star})\right\|^{2}\\
&+\frac{r(2-\gamma(r-1))}{(\gamma+2)^{2}}t^{\frac{2\gamma r}{\gamma+2}-2}\|X-X_{\star}\|^{2}\\
&+\int_{t_0}^{t}\frac{2r(2r+2-\gamma(r-1))(2-\gamma(r-1))}{(\gamma+2)^{3}}\\
&\qquad\qquad s^{\frac{2\gamma r}{\gamma+2}-3}\|X-X_{\star}\|^{2}ds\\
&+\int_{t_0}^{t}s^{\frac{2\gamma r}{\gamma+2}-1}\frac{2\gamma r}{\gamma+2}\\
&\qquad\left(f_{\star}-f(X)-\frac{1}{\gamma}\left\langle\nabla f(X),X_{\star}-X\right\rangle\right)ds.
\end{aligned}
$$

When $\gamma\geq1$ and $r\leq1+\frac{2}{\gamma}$, all terms are nonnegative, and we get

$$
f(X)-f_{\star}\leq\frac{E}{t^{\frac{2\gamma r}{\gamma+2}}},
$$

which recovers the result of (Aujol et al., 2019). Note that this rate is better than that of Section 3.2 since $\frac{2\gamma r}{\gamma+2} \geq \frac{2r}{3}$ for $\gamma \geq 1$.

3.4. SC-AGM

Wilson et al. (2021) presented the following ODE of the strongly convex accelerated gradient method (SC-AGM)

$$0 = \ddot{X} + 2\sqrt{\mu}\dot{X} + \nabla f(X) \tag{11}$$

with initial condition $X(0) = X_0$, $\dot{X}(0) = 0$, where $\mu > 0$ is the strong convexity parameter of f.

Consider the dilated coordinate $W = e^{\sqrt{\mu}t}(X - X_\star)$. The resulting conservation law with $t_0 = 0$ is

$$\begin{aligned} E &\equiv f(X_0) - f_\star \\ &= -\frac{\mu}{2}\left\|X_0 - X_\star\right\|^2 \\ &\quad + e^{\sqrt{\mu}t}\left(f(X) - f_\star + \frac{1}{2}\left\|\dot{X} + \sqrt{\mu}(X - X_\star)\right\|^2\right) \\ &\quad + \int_0^t \frac{\sqrt{\mu}e^{\sqrt{\mu}s}}{2}\left\|\dot{X}\right\|^2 ds + \int_0^t \sqrt{\mu}e^{\sqrt{\mu}s}(...)\, ds, \end{aligned}$$

where

$$\begin{aligned} (...) &= f_\star - f(X) - \langle \nabla f(X), X_\star - X\rangle - \frac{\mu}{2}\left\|X - X_\star\right\|^2 \\ &\geq 0. \end{aligned}$$

The inequality follows from μ-strong convexity of f. All the terms depending on t are nonnegative, thus $E + \frac{\mu}{2}\left\|X_0 - X_\star\right\|^2 \geq e^{\sqrt{\mu}t}(f(X) - f_\star)$, and we conclude

$$f(X) - f_\star \leq e^{-\sqrt{\mu}t}\left(f(X_0) - f_\star + \frac{\mu}{2}\left\|X_0 - X_\star\right\|^2\right).$$

This recovers the result of (Wilson et al., 2021).

3.5. Gradient flow

We conclude this section by showing that dilated coordinates also simplify the analysis of the gradient flow ODE

$$0 = \dot{X} + \nabla f(X)$$

with $X(0) = X_0$, which is a first-order ODE model of gradient descent.

Consider the dilated coordinate $W = t(X - X_\star)$. With $a(t) = 0$ in (8), we get the conservation law with $t_0 = 0$

$$\begin{aligned} E &\equiv -\frac{1}{2}\left\|X_0 - X_\star\right\|^2 \\ &= t\left(f(X) - f_\star\right) + \frac{1}{2}\left\|X - X_\star\right\|^2 - \left\|X_0 - X_\star\right\|^2 \\ &+ \int_0^t s\left\|\dot{X}\right\|^2 ds + \int_0^t \left(f_\star - f(X) - \langle \nabla f(X), X_\star - X\rangle\right)\, ds. \end{aligned}$$

We recover the well-known result

$$f(X) - f_\star \leq \frac{\left\|X_0 - X_\star\right\|^2}{2t}.$$

4. Continuous-time analysis of OGM-G

We now present a novel ODE model of OGM-G (Kim & Fessler, 2021), which optimally reduces the squared gradient magnitude (rather than the function value) for smooth convex minimization. Consider problem (1). Assume $f_\star = \inf_{x\in\mathbb{R}^n} f(x) > -\infty$. (We do not assume a solution exists.) Following steps similar to those of Su et al. (2014) with OGM-G, we obtain the OGM-G ODE

$$0 = \ddot{X} - \frac{3}{t - T}\dot{X} + 2\nabla f(X)$$

for $t \in (0, T)$ with initial value $X(0) = X_0$, $\dot{X}(0) = 0$. The precise derivation of the OGM-G ODE and the calculations throughout this section are presented in Appendix D.

Choose the dilated coordinate $W = (T - t)^\alpha(X - X_c)$ for some $X_c \in \mathbb{R}^n$. Since we expect the rate $\mathcal{O}\left(1/T^2\right)$, we choose $\alpha = -2$. The corresponding conservation law is

$$\begin{aligned} E &\equiv \frac{2}{T^2}(f(X_0) - f(X_c)) \\ &= \frac{2}{(T-t)^2}\left(f(X) - f(X_c)\right) - \frac{2}{(T-t)^4}\left\|X - X_c\right\|^2 \\ &+ \frac{1}{2(T-t)^4}\left\|(T-t)\dot{X} + 2(X - X_c)\right\|^2 \\ &+ \int_0^t \frac{4}{(T-s)^3}\left(f(X_c) - f(X) - \langle \nabla f(X), X_c - X\rangle\right) ds. \end{aligned}$$

4.1. OGM-G ODE $r = -3$

We now establish an $\mathcal{O}(1/T^2)$ rate on $\left\|\nabla f(X(T))\right\|^2$ via a conservation law. At first, this may seem curious as the conservation law contains no terms directly involving $\nabla f(X)$.

We first characterize the dynamics of the solution to the OGM-G ODE near the terminal time $t = T$.

Lemma 4.1. *Let $X\colon [0, T) \to \mathbb{R}^n$ be the solution to the OGM-G ODE. We can continuously extend $X(t)$, $\dot{X}(t)$, $\ddot{X}(t)$ to $t = T$ with*

$$\dot{X}(T) = 0, \quad \ddot{X}(T) = \lim_{t\to T^-}\frac{\dot{X}(t)}{t - T} = \nabla f(X(T)).$$

Proof outline. For simplicity, assume $\lim_{t\to T^-}\dot{X}(t)$ and $\lim_{t\to T^-}\ddot{X}(t) = \lim_{t\to T^-}\frac{\dot{X}(t) - \dot{X}(T)}{t - T}$ exist. We will formally prove these assumptions in Appendix D.3.

Consider the conservation law with $\alpha = 0$ and $X_c = X_0$:

$$E \equiv \frac{1}{2}\left\|\dot{X}\right\|^2 + 2(f(X) - f(X_0)) + \int_0^t \frac{3\left\|\dot{X}\right\|^2}{T - s}\, ds.$$

Since E is independent of time and since the first two terms are bounded, we have $\int_0^T \frac{3\left\|\dot{X}\right\|^2}{T-s} ds < \infty$. The finite integral implies $\lim_{t\to T^-} \dot{X}(t) = 0$. Furthermore,

$$\begin{aligned} 0 &= \lim_{t\to T^-} \left(\ddot{X}(t) - \frac{3}{t-T}\dot{X}(t) + 2\nabla f(X(t)) \right) \\ &= -2\ddot{X}(T) + 2\nabla f(X(T)). \end{aligned}$$ □

We now prove the promised result.

Theorem 4.2. *Let $X\colon [0,T] \to \mathbb{R}^n$ be the extended solution to the OGM-G ODE. Then X exhibits the rate*

$$\|\nabla f(X(T))\|^2 \le \frac{4\left(f(X_0) - f(X(T))\right)}{T^2} \le \frac{4\left(f(X_0) - f_\star\right)}{T^2}.$$

Proof. Consider the conservation law with $X_c = X(T)$ and define the Lyapunov function

$$\begin{aligned} \Phi(t) = &\frac{2}{(T-t)^2}\left(f(X) - f(X(T))\right) \\ &- \frac{2}{(T-t)^4}\left\|X - X(T)\right\|^2 \\ &+ \frac{1}{2(T-t)^4}\left\|(T-t)\dot{X} + 2(X - X(T))\right\|^2. \end{aligned}$$

Then $\Phi(t)$ is monotonically nonincreasing by the conservation law, and so $\Phi(0) \ge \lim_{t\to T^-}\Phi(t)$.

By applying L'Hôpital's rule,

$$\begin{aligned} \lim_{t\to T^-} \frac{f(X(t)) - f(X(T))}{(T-t)^2} &= \frac{1}{2}\left\|\nabla f(X(T))\right\|^2 \\ \lim_{t\to T^-} \frac{X(t) - X(T)}{(T-t)^2} &= \frac{1}{2}\nabla f(X(T)). \end{aligned}$$

Therefore,

$$\begin{aligned} \lim_{t\to T^-}\Phi(t) &= \left\|\nabla f(X(T))\right\|^2 - \frac{1}{2}\left\|\nabla f(X(T))\right\|^2 + 0 \\ &= \frac{1}{2}\left\|\nabla f(X(T))\right\|^2 \end{aligned}$$

and we conclude

$$\frac{1}{2}\left\|\nabla f(X(T))\right\|^2 \le \frac{2}{T^2}\left(f(X_0) - f(X(T))\right).$$ □

In the proof of Theorem 4.2, ∇f does not explicitly appear in the conservation law and only arises at the terminal time T due to Lemma 4.1. For this reason, we can establish a bound on $\left\|\nabla f(X(t))\right\|^2$ only at the terminal time.

Lee et al. (2021) presented the first Lyapunov analysis of the discrete-time OGM-G. We show in Appendix D.4 that the Lyapunov function of Theorem 4.2 is the continuous-time analog of the Lyapunov function of Lee et al. (2021). The discrete-time analysis for OGM-G also establish a rate on $\left\|\nabla f(x_k)\right\|^2$ only for the terminal iteration $k = K$.

4.2. OGM-G ODE for $r < -3$

Following Su et al. (2014), we generalize the OGM-G ODE to general r:

$$0 = \ddot{X} + \frac{r}{t-T}\dot{X} + 2\nabla f(X).$$

In Appendix D.3, we directly extend the arguments of Lemma 4.1 to conclude $\lim_{t\to T^-}\frac{\dot{X}(t)}{t-T} = -\frac{2}{r+1}\nabla f(X(T))$.

With the dilated coordinate $W = (T-t)^{-2}(X - X(T))$, we get the conservation law

$$\begin{aligned} E \equiv &\frac{2}{T^2}(f(X) - f(X(T))) + \frac{r+3}{T^4}\left\|X - X(T)\right\|^2 \\ = &\frac{2}{(T-t)^2}\left(f(X) - f(X(T))\right) + \frac{r+1}{(T-t)^4}\left\|X - X(T)\right\|^2 \\ &+ \frac{1}{2(T-t)^4}\left\|(T-t)\dot{X} + 2(X - X(T))\right\|^2 \\ &+ \int_0^t \frac{(-(r+3))}{(T-s)^5}\left\|(T-s)\dot{X} + 2(X - X(T))\right\|^2 ds \\ &+ \int_0^t \frac{4}{(T-s)^3}\left(f(X(T)) - f(X) - \langle\nabla f(X), X(T) - X\rangle\right) ds. \end{aligned}$$

Theorem 4.3. *Let $X\colon [0,T] \to \mathbb{R}^n$ be the extended solution to the OGM-G ODE with $r < -3$. Then,*

$$\left\|\nabla f(X(T))\right\|^2 \le \frac{2(-1-r)\left(f(X_0) - f(X(T))\right)}{T^2}$$

Proof outline. The arguments are similar to those of Theorem 4.2: Define a Lyapunov function $\Phi(t)$ based on the conservation law and consider the inequality $\Phi(0) \ge \lim_{t\to T^-}\Phi(t)$. Details are presented in Appendix D.5. □

4.3. Obtaining $\left\|\nabla f(X(T))\right\|^2 \le \mathcal{O}(1/T^4)$ with OGM + OGM-G ODE

We state a simple technique to obtain an $\mathcal{O}(\left\|x_0 - x_\star\right\|^2/T^4)$ rate from the $\mathcal{O}((f(X_0) - f_\star)/T^2)$ rate of the OGM-G ODE. This technique is based on the idea of Nesterov (2012), Nesterov et al. (2020) to concatenate AGM with OGM-G to obtain a $\left\|\nabla f(x_K)\right\|^2 \le \mathcal{O}(\left\|x_0 - x_\star\right\|^2/K^4)$ rate.

If one starts the AGM ODE with $X^{\mathrm{F}}(0) = X_0^{\mathrm{F}}$ and $\dot{X}^{\mathrm{F}}(0) = 0$, the terminal solution $X^{\mathrm{F}}(T)$ satisfies $f(X^{\mathrm{F}}(T)) - f_\star \le 2\left\|X_0 - X_\star\right\|^2/T^2$. Then we start the OGM-G ODE with $X^{\mathrm{G}}(0) = X^{\mathrm{F}}(T)$ and $\dot{X}^{\mathrm{G}}(0) = 0$ and obtain the solution $X^{\mathrm{G}}(T)$ satisfying $\left\|\nabla f(X^{\mathrm{G}}(T))\right\|^2 \le 4(f(X^{\mathrm{G}}(0)) - f_\star)/T^2$. Concatenating these two guarantees, we obtain $\left\|\nabla f(X^{\mathrm{G}}(T))\right\|^2 \le 8\left\|X_0 - X_\star\right\|^2/T^4$.

5. Discretization in dilated coordinates via semi-second-order symplectic Euler

In this section, we show that discretizing the AGM ODE ($r = 3$) using a semi-second-order symplectic Euler discretization in the dilated coordinate system leads to an algorithm with an $\mathcal{O}(1/k^2)$ rate. Despite the extensive prior work on continuous-time analyses and discretizations of the AGM ODE, obtaining an accelerated rate through a direct and "natural" discretization has been surprisingly tricky. Our result is the first to accomplish this, in the precise sense clarified in Section 5.1.

Again, the ODE (3), restated, is $0 = \ddot{X} + \frac{3}{t}\dot{X} + \nabla f(X)$. With $W = t^2(X - X_\star)$, the ODE (6), restated, is

$$0 = \frac{1}{t^2}\ddot{W} - \frac{1}{t^3}\dot{W} + \nabla_W U(W,t). \tag{6}$$

We first identify a generalized coordinate W and conjugate momentum P to replace X and $\dot{X}$. The dilated coordinate $W = t^2(X - X_\star)$ has been chosen, so we determine the generalized momentum via the Lagrangian formulation.

Recall from (5) that $U(W,t) = t^2\left(f(X(W,t)) - f_\star\right)$. Define the Lagrangian as

$$L(W,\dot{W},t) = \frac{1}{2t}\left\|\dot{W}\right\|^2 - t\,U(W,t).$$

Then the Euler–Lagrange equation $\frac{d}{dt}\nabla_{\dot{W}}L = \nabla_W L$ yields the ODE (6) and $P = \nabla_{\dot{W}}L = \frac{\dot{W}}{t} = t\dot{X} + 2(X - X_\star)$ is the conjugate momentum. Express (6) in W and P:

$$\dot{P} = -t\nabla f(X(W,t))$$
$$\dot{W} = tP$$

and $\ddot{W} = P - t^2\,\nabla f\left(X(W,t)\right)$.

Inspired by the symplectic Euler (Hairer et al., 2006) and velocity Verlet integrators (Verlet, 1968; Swope et al., 1982; Allen & Tildesley, 2017) we consider alternating updates of W and P but use a second-order update for W:

$$P(t+h) \approx P(t) - t\nabla f(X)h$$
$$W(t+h) \approx W(t) + \dot{W}(t)h + \ddot{W}(t)\frac{h^2}{2}$$
$$= W(t) + tP(t)h + \left(P(t) - t^2\nabla f(X(W,t))\right)\frac{h^2}{2}.$$

We refer to this method as a semi-second-order symplectic Euler. This discretization is also an instance of the Nyström method (Hairer et al., 2006).

Identifying w_k and p_k with $W(hk)$ and $P(hk)$ and defining x_k through $w_k = h^2k^2(x_k - X_\star)$, we get the method

$$p_{k+1} = p_k - kh^2\nabla f\left(x_k\right)$$
$$x_{k+1} = \frac{k^2}{(k+1)^2}\left(x_k - \frac{h^2}{2}\nabla f\left(x_k\right)\right) + \frac{2k+1}{(k+1)^2}\left(\frac{p_{k+1}}{2} + X_\star\right).$$

Finally, letting $s = h^2$, $\theta_k = \frac{k}{2}$ and $z_k = \frac{p_k}{2} + X_\star$, we get

$$\begin{aligned} x_k^+ &= x_k - \frac{s}{2}\nabla f(x_k) \\ z_{k+1} &= z_k - s\theta_k\nabla f(x_k) \\ x_{k+1} &= \frac{\theta_k^2}{\theta_{k+1}^2}x_k^+ + \left(1 - \frac{\theta_k^2}{\theta_{k+1}^2}\right)z_{k+1} \end{aligned} \tag{12}$$

for $k = 0, 1, \ldots$. The starting point is $x_0 = z_0 = X_0 \in \mathbb{R}^n$, since z_0 corresponds to $\frac{P(0)}{2} + X_\star = X_0$.

Theorem 5.1. *Assume f is convex and L-smooth. Assume f has a minimizer $X_\star$. For $s \in \left(0, \frac{2}{L}\right]$, (12) exhibits the rate*

$$f(x_k^+) - f_\star \le \frac{2\left\|X_0 - X_\star\right\|^2}{sk^2}.$$

Proof outline. The proof is based on the Lyapunov analysis $\Phi_k \le \Phi_{k-1} \le \cdots \le \Phi_0$ with

$$\Phi_k = 2c_k\theta_k^2\left(f(x_k) - f_\star - \frac{s}{4}\left\|\nabla f(x_k)\right\|^2\right) + \frac{1}{s}\left\|z_{k+1} - X_\star\right\|^2$$

and $c_k = \frac{\theta_{k+1}}{\theta_{k+1}^2 - \theta_k^2}$ for $k = 0, 1, \ldots$. The details are presented in Appendix E. □

5.1. Discussion

Hamiltonian mechanics. Some may wonder what can be said from a Hamiltonian mechanics perspective. We discuss this matter briefly in Appendix F, and (Diakonikolas & Jordan, 2021; França et al., 2021a) pursues this direction deeply. Here, we point out the quick observation that the explicit time-dependence of the Lagrangian makes the Hamiltonian time-dependent, and this time-dependence makes the Hamiltonian a non-conserved quantity. Therefore, the classical theory of symplectic integrators is not immediately applicable, but we nevertheless use our method and obtain an accelerated rate.

Prior discretizations. The discretization of (Wibisono et al., 2016) achieves an $\mathcal{O}(1/k^2)$ rate, but, arguably, this discretization "does not flow natural from the dynamical-systems framework" (Jordan, 2018, p. 529). Zhang et al. (2018) achieved an accelerated rate with a Runge–Kutta method, but their $\mathcal{O}(1/k^2)$ rate requires the additional assumption of infinite differentiability. Shi et al. (2019) used a symplectic integrator with $\dot{X}$ as the momentum (no coordinate change) and achieved an $\mathcal{O}(1/k^2)$ rate, but they crucially rely on the high-resolution ODE formulation. França et al. (2021a) proposed a generalized symplectic integrator and established $\mathcal{O}(1/k^2)$ rate for exponentially large k depending on the stepsize, but their rate does not hold for all $k \in \mathbb{N}$. Even et al. (2021) introduced alternative "continuized" framework and obtained $\mathcal{O}(1/k^2)$ with randomized discretizations. On the other hand, our result is a direct, non-randomized discretization of the AGM ODE

that achieves an $\mathcal{O}(1/k^2)$ rate without making additional assumptions or using a high-resolution formulation.

Discretized rate surpasses AGM. The rate of Theorem 5.1 with $s = \frac{2}{L}$ is

$$f(x_k^+) - f_\star \leq \frac{L\left\|X_0 - X_\star\right\|^2}{k^2}.$$

Interestingly, this rate is smaller (better) than the rate of Nesterov's AGM by a factor of 2 (Nesterov, 1983) but is slightly larger (worse) than the exact optimal rate of OGM (Drori & Teboulle, 2014; Kim & Fessler, 2016; Drori, 2017). This improvement seems to be in part due to the choice of Lyapunov function, inspired by (Park et al., 2021), that allows a tighter analysis. By taking the continuous-time limit of AGM and then discretizing, we arrived at a discretized algorithm that is *better* than the original AGM.

Interpreting z_k as conjugate momentum. Lee et al. (2021) point out that many known accelerated gradient methods have an auxiliary z_k-sequence satisfying a geometric structure. In our analysis of the AGM ODE, we identify that z_k is (up to a factor-2 scaling and translation with $X_\star$) the conjugate momentum $P = \dot{W}/t = t\dot{X} + 2(X - X_\star)$ of the dilated coordinate $W = t^2(X - X_\star)$.

Moreover, we've observed that this interpretation of the z-variables as conjugate momenta of the dilated coordinate systems (with some rescaling and translation) also holds in other setups, including the SC-AGM and the OGM-G setups. Specifically, when we discretize the ODEs in the dilated coordinate systems $W(t)$, the discretized methods closely resemble the known accelerated methods, and the z-variables roughly correspond to conjugate momenta $P(t)$. We leave the formalization and development of this observation as future work.

6. Conclusion

This work presents a methodology for analyzing continuous-time models of accelerated gradient methods through deriving conservation laws in dilated coordinate systems. Using this methodology, we recover many known continuous-time analyses in a streamlined manner and obtain novel continuous-time analyses of OGM-G.

We hypothesize that our dilated coordinates can simplify analyses of other setups beyond those explored in Sections 3 and 4. For example, exploring the use of dilated coordinates in stochastic differential equations modeling stochastic optimization and investigating whether dilated coordinates generally simplify discretization, as was the case for the AGM ODE ($r = 3$) in Section 5, are interesting directions of future work. Finally, finding a more fundamental understanding of the interpretation of z_k as the conjugate momentum would also be interesting.

Acknowledgements

JJS and EKR were supported by the Samsung Science and Technology Foundation grant (Project Number SSTF-BA2101-02). We thank Jongmin Lee for valuable discussions about OGM-G. We thank Chanwoo Park for reviewing the manuscript and providing valuable feedback. Finally, we thank the anonymous reviewers for their thoughtful comments.

References

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

Adly, S. and Attouch, H. Finite time stabilization of continuous inertial dynamics combining dry friction with Hessian-driven damping. *HAL-02557928*, 2020b.

Adly, S. and Attouch, H. First-order inertial algorithms involving dry friction damping. *Mathematical Programming*, 2021.

Adly, S., Attouch, H., and Le, M. H. First order inertial optimization algorithms with threshold effects associated with dry friction. *HAL-03284220*, 2021a.

Adly, S., Attouch, H., and Vo, V. N. Asymptotic behavior of Newton-like inertial dynamics involving the sum of potential and nonpotential terms. *Fixed Point Theory and Algorithms for Sciences and Engineering*, 2021(1):17, 2021b.

Adly, S., Attouch, H., and Vo, V. N. Newton-type inertial algorithms for solving monotone equations governed by sums of potential and nonpotential operators. *HAL-03260201*, 2021c.

Allen, M. P. and Tildesley, D. J. *Computer Simulation of Liquids*. Oxford University Press, second edition, 2017.

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

Alvarez, F., Attouch, H., Bolte, J., and 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.

Apidopoulos, V., Aujol, J.-F., and Dossal, C. H. On a second order differential inclusion modeling the FISTA algorithm. *HAL-01517708*, 2017.

Apidopoulos, V., Aujol, J.-F., and Dossal, C. The differential inclusion modeling FISTA algorithm and optimality of convergence rate in the case $b \leq 3$. *SIAM Journal on Optimization*, 28(1):551–574, 2018.

Apidopoulos, V., Aujol, J.-F., Dossal, C., and Rondepierre, A. Convergence rates of an inertial gradient descent algorithm under growth and flatness conditions. *Mathematical Programming*, 187(1):151–193, 2021.

Attouch, H. and Alvarez, F. The heavy ball with friction dynamical system for convex constrained minimization problems. *Belgian-French-German Conference on Optimization*, 1998.

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

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

Attouch, H. and Cabot, A. Convergence rates of inertial forward-backward algorithms. *SIAM Journal on Optimization*, 28(1):849–874, 2018b.

Attouch, H. and Cabot, A. Convergence rate of a relaxed inertial proximal algorithm for convex minimization. *Optimization*, 69(6):1281–1312, 2020.

Attouch, H. and Chbani, Z. Fast inertial dynamics and FISTA algorithms in convex optimization. Perturbation aspects. *arXiv:1507.01367*, 2015.

Attouch, H. and Czarnecki, M.-O. Asymptotic control and stabilization of nonlinear oscillators with non-isolated equilibria. *Journal of Differential Equations*, 179(1):278–310, 2002.

Attouch, H. and Czarnecki, M.-O. Asymptotic behavior of gradient-like dynamical systems involving inertia and multiscale aspects. *Journal of Differential Equations*, 262 (3):2745–2770, 2017.

Attouch, H. and Laszlo, S. Convex optimization via inertial algorithms with vanishing Tikhonov regularization: Fast convergence to the minimum norm solution. *arXiv:2104.11987*, 2021.

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

Attouch, H. and László, S. C. Continuous Newton-like inertial dynamics for monotone inclusions. *Set-Valued and Variational Analysis*, 29(3):555–581, 2021.

Attouch, H. and Maingé, P.-É. Asymptotic behavior of second-order dissipative evolution equations combining potential with non-potential effects. *ESAIM: Control, Optimisation and Calculus of Variations*, 17(3):836–857, 2011.

Attouch, H. and Peypouquet, J. The rate of convergence of Nesterov's accelerated forward-backward method is actually faster than $1/k^2$. *SIAM Journal on Optimization*, 26(3):1824–1834, 2016.

Attouch, H. and Peypouquet, J. Convergence of inertial dynamics and proximal algorithms governed by maximally monotone operators. *Mathematical Programming*, 174 (1):391–432, 2019.

Attouch, H., Bolte, J., and Redont, P. Optimizing properties of an inertial dynamical system with geometric damping: Link with proximal methods. *Control and Cybernetics*, 31(3):643–657, 2002.

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

Attouch, H., Cabot, A., Chbani, Z., and Riahi, H. Inertial forward–backward algorithms with perturbations: Application to Tikhonov regularization. *Journal of Optimization Theory and Applications*, 179(1):1–36, 2018a.

Attouch, H., Cabot, A., Chbani, Z., and Riahi, H. Rate of convergence of inertial gradient dynamics with time-dependent viscous damping coefficient. *Evolution Equations & Control Theory*, 7(3):353–371, 2018b.

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

Attouch, H., Chbani, Z., and Riahi, H. Combining fast inertial dynamics for convex optimization with Tikhonov regularization. *Journal of Mathematical Analysis and Applications*, 457(2):1065–1094, 2018d.

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

Attouch, H., Chbani, Z., and Riahi, H. Fast convex optimization via time scaling of damped inertial gradient dynamics. *HAL-02138954*, 2019b.

Attouch, H., Chbani, Z., and Riahi, H. Rate of convergence of the Nesterov accelerated gradient method in the subcritical case $\alpha \leq 3$. *ESAIM: Control, Optimisation and Calculus of Variations*, 25:2, 2019c.

Attouch, H., Chbani, Z., Fadili, J., and Riahi, H. First-order optimization algorithms via inertial systems with Hessian driven damping. *Mathematical Programming*, 2020a.

Attouch, H., Chbani, Z., and Riahi, H. Convergence rate of inertial proximal algorithms with general extrapolation and proximal coefficients. *Vietnam Journal of Mathematics*, 48(2):247–276, 2020b.

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

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

Attouch, H., Chbani, Z., Fadili, J., and Riahi, H. Fast convergence of dynamical ADMM via time scaling of damped inertial dynamics. *Journal of Optimization Theory and Applications*, 2021c.

Attouch, H., Fadili, J., and Kungurtsev, V. On the effect of perturbations, errors in first-order optimization methods with inertia and Hessian driven damping. *arXiv:2106.16159*, 2021d.

Attouch, H., Balhag, A., Chbani, Z., and Riahi, H. Damped inertial dynamics with vanishing Tikhonov regularization: Strong asymptotic convergence towards the minimum norm solution. *Journal of Differential Equations*, 311: 29–58, 2022.

Aujol, J.-F. and Dossal, C. Optimal rate of convergence of an ODE associated to the fast gradient descent schemes for $b > 0$. *HAL-01547251*, 2017a.

Aujol, J.-F. and Dossal, C. H. The optimal decay for the solution of the monotone inclusion associated to FISTA for $b \leq 3$ is $2b/3$. *HAL-01565933*, 2017b.

Aujol, J.-F., Dossal, C., and Rondepierre, A. Optimal convergence rates for Nesterov acceleration. *SIAM Journal on Optimization*, 29(4):3131–3153, 2019.

Aujol, J.-F., Dossal, C., and Rondepierre, A. Convergence rates of the Heavy-Ball method for quasi-strongly convex optimization. *HAL-02545245*, 2021.

Aujol, J.-F., Dossal, C., and Rondepierre, A. Convergence rates of the Heavy-Ball method under the Łojasiewicz property. *Mathematical Programming*, 2022.

Bauschke, H. H. and Combettes, P. L. *Convex Analysis and Monotone Operator Theory in Hilbert Spaces*. Springer International Publishing, second edition, 2017.

Betancourt, M., Jordan, M. I., and Wilson, A. C. On symplectic optimization. *arXiv:1802.03653*, 2018.

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

Boţ, R. I. and Csetnek, E. R. Second-order dynamical systems associated to variational inequalities. *Applicable Analysis*, 96(5):799–809, 2017.

Boţ, R. I. and Csetnek, E. R. Convergence rates for forward–backward dynamical systems associated with strongly monotone inclusions. *Journal of Mathematical Analysis and Applications*, 457(2):1135–1152, 2018.

Boţ, R. I. and Csetnek, E. R. A second-order dynamical system with Hessian-driven damping and penalty term associated to variational inequalities. *Optimization*, 68 (7):1265–1277, 2019.

Bot, R. I. and Hulett, D. A. Second order splitting dynamics with vanishing damping for additively structured monotone inclusions. *arXiv:2201.01017*, 2022.

Boţ, R. I. and Nguyen, D.-K. Improved convergence rates and trajectory convergence for primal-dual dynamical systems with vanishing damping. *Journal of Differential Equations*, 303:369–406, 2021.

Boţ, R. I., Csetnek, E. R., and László, S. C. Second-order dynamical systems with penalty terms associated to monotone inclusions. *Analysis and Applications*, 16(05):601–622, 2018.

Boţ, R. I., Csetnek, E. R., and László, S. C. Tikhonov regularization of a second order dynamical system with Hessian driven damping. *Mathematical Programming*, 189(1):151–186, 2021.

Boyd, S. and Vandenberghe, L. *Convex Optimization*. Cambridge University Press, 2004.

Chambolle, A. and Dossal, C. On the convergence of the iterates of the "fast iterative shrinkage/thresholding algorithm". *Journal of Optimization Theory and Applications*, 166(3):968–982, 2015.

d'Aspremont, A., Scieur, D., and Taylor, A. Acceleration methods. *Foundations and Trends® in Optimization*, 5 (1–2):1–245, 2021.

Diakonikolas, J. and Jordan, M. I. Generalized momentum-based methods: A Hamiltonian perspective. *SIAM Journal on Optimization*, 31(1):915–944, 2021.

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

Drori, Y. The exact information-based complexity of smooth convex minimization. *Journal of Complexity*, 39:1–16, 2017.

Drori, Y. and Teboulle, M. Performance of first-order methods for smooth convex minimization: A novel approach. *Mathematical Programming*, 145(1):451–482, 2014.

Even, M., Berthier, R., Bach, F., Flammarion, N., Hendrikx, H., Gaillard, P., Massoulié, L., and Taylor, A. Continuized accelerations of deterministic and stochastic gradient descents, and of gossip algorithms. *Neural Information Processing Systems*, 2021.

França, G., Robinson, D., and Vidal, R. ADMM and accelerated ADMM as continuous dynamical systems. *International Conference on Machine Learning*, 2018.

França, G., Sulam, J., Robinson, D., and Vidal, R. Conformal symplectic and relativistic optimization. *Neural Information Processing Systems*, 2020a.

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

França, G., Jordan, M. I., and Vidal, R. On dissipative symplectic integration with applications to gradient-based optimization. *Journal of Statistical Mechanics: Theory and Experiment*, 2021(4):043402, 2021a.

França, G., Robinson, D. P., and Vidal, R. Gradient flows and proximal splitting methods: A unified view on accelerated and stochastic optimization. *Physical Review E: Statistical Physics, Plasmas, Fluids, and Related Interdisciplinary Topics*, 103(5):053304, 2021b.

Hairer, E., Lubich, C., and Gerhard, W. *Geometric Numerical Integration: Structure-Preserving Algorithms for Ordinary Differential Equations*. Springer, second edition, 2006.

Hassan-Moghaddam, S. and Jovanović, M. R. Proximal gradient flow and Douglas–Rachford splitting dynamics: Global exponential stability via integral quadratic constraints. *Automatica*, 123:109311, 2021.

Hu, B. and Lessard, L. Dissipativity theory for Nesterov's accelerated method. *International Conference on Machine Learning*, 2017.

Jordan, M. I. Dynamical, symplectic and stochastic perspectives on gradient-based optimization. In *International Congress of Mathematicians*, pp. 523–549. World Scientific, 2018.

Kim, D. and Fessler, J. A. Optimized first-order methods for smooth convex minimization. *Mathematical Programming*, 159(1-2):81–107, 2016.

Kim, D. and Fessler, J. A. Optimizing the efficiency of first-order methods for decreasing the gradient of smooth convex functions. *Journal of Optimization Theory and Applications*, 188(1):192–219, 2021.

Krichene, W., Bayen, A., and Bartlett, P. L. Accelerated mirror descent in continuous and discrete time. *Neural Information Processing Systems*, 2015.

Lee, J., Park, C., and Ryu, E. K. A geometric structure of acceleration and its role in making gradients small fast. *Neural Information Processing Systems*, 2021.

Maddison, C. J., Paulin, D., Teh, Y. W., O'Donoghue, B., and Doucet, A. Hamiltonian descent methods. *arXiv:1809.05042*, 2018.

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

Muehlebach, M. and Jordan, M. A dynamical systems perspective on Nesterov acceleration. *International Conference on Machine Learning*, 2019.

Muehlebach, M. and Jordan, M. Continuous-time lower bounds for gradient-based algorithms. *International Conference on Machine Learning*, 2020.

Muehlebach, M. and Jordan, M. I. Optimization with momentum: Dynamical, control-theoretic, and symplectic perspectives. *Journal of Machine Learning Research*, 22 (73):1–50, 2021.

Nemirovski, A. *Optimization II. Numerical Methods for Nonlinear Continuous Optimization*. Lecture Note, The Israel Institute of Technology Faculty of Industrial Engineering and Management, 1999.

Nemirovsky, A. On optimality of Krylov's information when solving linear operator equations. *Journal of Complexity*, 7(2):121–130, 1991.

Nemirovsky, A. S. Information-based complexity of linear operator equations. *Journal of Complexity*, 8(2):153–175, 1992.

Nesterov, Y. A method of solving a convex programming problem with convergence rate $o(1/k^2)$. *Doklady Akademii Nauk SSSR*, 269(3):543–547, 1983.

Nesterov, Y. *Introductory Lectures on Convex Optimization: A Basic Course*. Springer, 2004.

Nesterov, Y. How to make the gradients small. *Optima. Mathematical Optimization Society Newsletter*, (88):10–11, 2012.

Nesterov, Y. *Lectures on Convex Optimization*. Springer, second edition, 2018.

Nesterov, Y., Gasnikov, A., Guminov, S., and Dvurechensky, P. Primal-dual accelerated gradient methods with small-dimensional relaxation oracle. *Optimization Methods and Software*, 2020.

Park, C. and Ryu, E. K. Optimal first-order algorithms as a function of inequalities. *arXiv:2110.11035*, 2021.

Park, C., Park, J., and Ryu, E. K. Factor-$\sqrt{2}$ acceleration of accelerated gradient methods. *arXiv:2102.07366*, 2021.

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

Rockafellar, R. T. *Convex Analysis*. Princeton University Press, 1970.

Ryu, E. K. and Yin, W. *Large-Scale Convex Optimization via Monotone Operators*. Cambridge University Press, 2022.

Scieur, D., Roulet, V., Bach, F., and d'Aspremont, A. Integration methods and optimization algorithms. *Neural Information Processing Systems*, 2017.

Sebbouh, O., Dossal, C., and Rondepierre, A. Nesterov's acceleration and Polyak's heavy ball method in continuous time: Convergence rate analysis under geometric conditions and perturbations. *arXiv:1907.02710*, 2019.

Shi, B., Du, S. S., Su, W., and Jordan, M. I. Acceleration via symplectic discretization of high-resolution differential equations. *Neural Information Processing Systems*, 2019.

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

Su, W., Boyd, S., and Candès, E. J. A differential equation for modeling Nesterov's accelerated gradient method: Theory and insights. *Neural Information Processing Systems*, 2014.

Su, W., Boyd, S., and Candès, E. J. A differential equation for modeling Nesterov's accelerated gradient method: Theory and insights. *Journal of Machine Learning Research*, 17(153):1–43, 2016.

Swope, W. C., Andersen, H. C., Berens, P. H., and Wilson, K. R. A computer simulation method for the calculation of equilibrium constants for the formation of physical clusters of molecules: Application to small water clusters. *The Journal of Chemical Physics*, 76(1):637–649, 1982.

Verlet, L. Computer "experiments" on classical fluids. II. Equilibrium correlation functions. *Physical Review*, 165 (1):201–214, 1968.

Wibisono, A., Wilson, A. C., and Jordan, M. I. A variational perspective on accelerated methods in optimization. *Proceedings of the National Academy of Sciences*, 113(47): E7351–E7358, 2016.

Wilson, A. C., Mackey, L., and Wibisono, A. Accelerating rescaled gradient descent: Fast optimization of smooth functions. *Neural Information Processing Systems*, 2019.

Wilson, A. C., Recht, B., and Jordan, M. I. A Lyapunov analysis of accelerated methods in optimization. *Journal of Machine Learning Research*, 22(113):1–34, 2021.

Zhang, J., Mokhtari, A., Sra, S., and Jadbabaie, A. Direct Runge–Kutta discretization achieves acceleration. *Neural Information Processing Systems*, 2018.

Zhang, J., Sra, S., and Jadbabaie, A. Acceleration in first order quasi-strongly convex optimization by ODE discretization. *Conference on Decision and Control*, 2019.

Zhang, P., Orvieto, A., and Daneshmand, H. Rethinking the variational interpretation of accelerated optimization methods. *Neural Information Processing Systems*, 2021.

Zhou, K., Tian, L., So, A. M.-C., and Cheng, J. Practical schemes for finding near-stationary points of convex finite-sums. *International Conference on Artificial Intelligence and Statistics*, 2022.

A. Partial derivative notation

For $U\colon \mathbb{R}^n \times \mathbb{R} \to \mathbb{R}$, we assign symbols $W \in \mathbb{R}^n$ and $t \in \mathbb{R}$ for the inputs, i.e., we write $U(W,t)$. At the same time, we consider the curve $W\colon \mathbb{R} \to \mathbb{R}^n$ a function of $t \in \mathbb{R}$, i.e., we write $W(t)$. When we provide the curve $W(t)$ as the first input to U, we get $U(W(t),t)$, which is now a function solely of $t \in \mathbb{R}$, and we can take the total derivative $\frac{d}{dt}$ of it. Using the chain rule of vector calculus, we get

$$\frac{d}{dt}U(W(t),t) = \left\langle (D_1U)(W(t),t), \dot{W}(t) \right\rangle + (D_2U)(W(t),t)$$

where D_1U is the derivative of $U(\cdot,\cdot)$ with respect to the first n coordinates and D_2U is the derivative of $U(\cdot,\cdot)$ with respect to the last coordinate. When $U(W,t)$ is viewed as a function of W and t (when W is an input variable independent of t rather than a curve), then

$$D_1U = \nabla_W U, \qquad D_2U = \frac{\partial}{\partial t}U.$$

Write $\nabla_W U(W(t),t)$ to mean take the partial derivative of $U(W,t)$ with respect to W and then plug in $W = W(t)$. Likewise, write $\frac{\partial}{\partial t}U(W(t),t)$ to mean take the partial derivative of $U(W,t)$ with respect to t and then plug in $W = W(t)$. Finally, we can write

$$\begin{aligned}\frac{d}{dt}U(W(t),t) &= \left\langle (D_1U)(W(t),t), \dot{W}(t) \right\rangle + (D_2U)(W(t),t) \\ &= \left\langle \nabla_W U(W(t),t), \dot{W}(t) \right\rangle + \frac{\partial}{\partial t}U(W(t),t).\end{aligned}$$

B. Comparison with (Diakonikolas & Jordan, 2021)

Diakonikolas & Jordan (2021) present a methodology based on Hamiltonian mechanics, and their goal is also to provide a unified methodology for analyzing continuous-time models of accelerated gradient methods. However, our methodology differs from that of Diakonikolas & Jordan (2021) in the following three ways.

- We start from a given ODE and derive conservations laws, while Diakonikolas & Jordan (2021) start from a Hamiltonian and derive the ODE.
- In our framework, different choices of 'α' produce different conservation laws for one fixed ODE, but in (Diakonikolas & Jordan, 2021) different choices of 'α' corresponds to different ODEs and different corresponding energies.
- Our framework accommodates translation with respect to an arbitrary "center point" X_c.

Our analyses of the AGM, SC-AGM, and OGM-G ODEs crucially rely on these differences and therefore cannot be obtained by the methodology of Diakonikolas & Jordan (2021) as-is:

- The approach of Diakonikolas & Jordan (2021) does not lead to a Lyapunov function or a conservation law containing $\|W\|^2$. Many of our results crucially rely on using an energy $U(W,t)$ with the $\|W\|^2$ term.
- The translation with respect to $X_c = X(T)$ is essential for the analysis of OGM-G ODE in Theorem 4.2.

C. Omitted calculations of Section 3

C.1. Conservation law for generalized r

We start with ODE (9)

$$0 = \ddot{X} + \frac{r}{t}\dot{X} + \nabla f(X).$$

Now consider the coordinate change $W = t^\alpha(X - X_\star)$. Then we see

$$\begin{aligned}W &= t^\alpha(X - X_\star) \\ \dot{W} &= t^\alpha \dot{X} + \alpha t^{\alpha-1}(X - X_\star) \\ \ddot{W} &= t^\alpha \ddot{X} + 2\alpha t^{\alpha-1}\dot{X} + \alpha(\alpha-1)t^{\alpha-2}(X - X_\star).\end{aligned}$$

From this, we can rewrite X, $\dot{X}$, $\ddot{X}$ in terms of W, $\dot{W}$, $\ddot{W}$,

$$\begin{aligned}
X &= \frac{W}{t^\alpha} + X_\star \\
\dot{X} &= \frac{\dot{W}}{t^\alpha} - \alpha\frac{W}{t^{\alpha+1}} \\
\ddot{X} &= \frac{1}{t^\alpha}\ddot{W} - \frac{2\alpha}{t^{\alpha+1}}\dot{W} + \frac{\alpha(\alpha+1)}{t^{\alpha+2}}W.
\end{aligned}$$

Plugging these to (9) we get ODE

$$0 = \frac{1}{t^\alpha}\ddot{W} + \frac{r-2\alpha}{t^{\alpha+1}}\dot{W} + \frac{\alpha(\alpha+1-r)}{t^{\alpha+2}}W + \nabla f\left(\frac{W}{t^\alpha} + X_\star\right).$$

Now by defining

$$U(W,t) = \frac{\alpha(\alpha+1-r)}{2t^{\alpha+2}}\|W\|^2 + t^\alpha\left(f\left(\frac{W}{t^\alpha}+X_\star\right) - f_\star\right)$$

we can rewrite the ODE as

$$0 = \frac{1}{t^\alpha}\ddot{W} + \frac{r-2\alpha}{t^{\alpha+1}}\dot{W} + \nabla_W U(W,t). \tag{13}$$

Now plugging $a(t) = \frac{1}{t^\alpha}$, $b(t) = \frac{r-2\alpha}{t^{\alpha+1}}$, from conservation law (8) we get

$$\begin{aligned}
E &\equiv \frac{1}{2t_0^\alpha}\left\|\dot{W}(t_0)\right\|^2 + \frac{\alpha(\alpha+1-r)}{2t_0^{\alpha+2}}\|W(t_0)\|^2 + t_0^\alpha\left(f\left(\frac{W(t_0)}{t_0^\alpha}+X_\star\right) - f_\star\right) \\
&= \frac{1}{2t^\alpha}\left\|\dot{W}\right\|^2 + \frac{\alpha(\alpha+1-r)}{2t^{\alpha+2}}\|W\|^2 + t^\alpha\left(f\left(\frac{W}{t^\alpha}+X_\star\right) - f_\star\right) + \int_{t_0}^t \frac{2r-3\alpha}{2s^{\alpha+1}}\left\|\dot{W}\right\|^2 ds \\
&\quad - \int_{t_0}^t \left(\alpha s^{\alpha-1}\left(f\left(\frac{W}{s^\alpha}+X_\star\right) - f_\star - \left\langle\nabla f\left(\frac{W}{s^\alpha}+X_\star\right), \frac{W}{s^\alpha}\right\rangle\right) - \frac{\alpha(\alpha+1-r)(\alpha+2)}{2s^{\alpha+3}}\|W\|^2\right) ds.
\end{aligned}$$

Rewriting in terms of X, $\dot{X}$, $\ddot{X}$ with some reordering we have

$$\begin{aligned}
E &\equiv t_0^\alpha\left(f\left(X(t_0)\right) - f_\star\right) + \frac{1}{2}t_0^{\alpha-2}\left\|t_0\dot{X}(t_0) + \alpha(X(t_0)-X_\star)\right\|^2 + \frac{\alpha(\alpha+1-r)}{2}t_0^{\alpha-2}\|X(t_0)-X_\star\|^2 \qquad (10) \\
&= t^\alpha\left(f\left(X\right) - f_\star\right) + \frac{1}{2}t^{\alpha-2}\left\|t\dot{X} + \alpha(X-X_\star)\right\|^2 + \frac{\alpha(\alpha+1-r)}{2}t^{\alpha-2}\|X-X_\star\|^2 \\
&\quad + \int_{t_0}^t\left(\frac{(2r-3\alpha)s^{\alpha-3}}{2}\left\|s\dot{X} + \alpha(X-X_\star)\right\|^2 + \frac{\alpha(\alpha+1-r)(\alpha+2)}{2}s^{\alpha-3}\|X-X_\star\|^2\right)ds \\
&\quad + \int_{t_0}^t \alpha s^{\alpha-1}\left(f_\star - f(X) - \langle\nabla f(X), X_\star - X\rangle\right)ds.
\end{aligned}$$

C.2. AGM ODE with $r > 3$

Plugging $\alpha = 2$, $t_0 = 0$ to (10), we have

$$\begin{aligned}
E &\equiv (5-r)\|X_0 - X_\star\|^2 \\
&= t^2\left(f\left(X\right) - f_\star\right) + \frac{1}{2}\left\|t\dot{X} + 2(X-X_\star)\right\|^2 + (3-r)\|X-X_\star\|^2 \\
&\quad + \int_0^t\left(\frac{r-3}{s}\left\|s\dot{X} + 2(X-X_\star)\right\|^2 + \frac{4(3-r)}{s}\|X-X_\star\|^2\right)ds \\
&\quad + \int_0^t 2s\left(f_\star - f(X) - \langle\nabla f(X), X_\star - X\rangle\right)ds.
\end{aligned}$$

Also, since

$$\begin{aligned}
&\int_0^t \left(\frac{r-3}{s} \left\| s\dot{X} + 2(X - X_\star) \right\|^2 + \frac{4(3-r)}{s} \|X - X_\star\|^2 \right) ds \\
&= \int_0^t \left(\frac{r-3}{s} \left\| s\dot{X} \right\|^2 + 4(r-3) \left\langle \dot{X}, X - X_\star \right\rangle \right) ds = \int_0^t \frac{r-3}{s} \left\| s\dot{X} \right\|^2 ds + \left[2(r-3) \|X - X_\star\|^2 \right]_0^t \\
&= \int_0^t \frac{r-3}{s} \left\| s\dot{X} \right\|^2 ds + 2(r-3) \left(\|X - X_\star\|^2 - \|X_0 - X_\star\|^2 \right).
\end{aligned}$$

Therefore

$$\begin{aligned}
E &\equiv (5-r) \|X_0 - X_\star\|^2 \\
&= t^2 (f(X) - f_\star) + \frac{1}{2} \left\| t\dot{X} + 2(X - X_\star) \right\|^2 + (r-3) \|X - X_\star\|^2 - 2(r-3) \|X_0 - X_\star\|^2 \\
&\quad + \int_0^t \frac{r-3}{s} \left\| s\dot{X} \right\|^2 ds + \int_0^t 2s \big(f_\star - f(X) - \langle \nabla f(X), X_\star - X \rangle \big) ds.
\end{aligned}$$

C.3. AGM ODE with growth condition

Rescaling (13) by multiplying t^β we get

$$0 = \frac{1}{t^{\alpha-\beta}} \ddot{W} + \frac{r - 2\alpha}{t^{\alpha-\beta+1}} \dot{W} + \nabla_W \left(\frac{\alpha(\alpha+1-r)}{2t^{\alpha-\beta+2}} \|W\|^2 + t^{\alpha+\beta} \left(f\left(\frac{W}{t^\alpha} + X_\star \right) - f_\star \right) \right).$$

Now plugging $a(t) = \frac{1}{t^{\alpha-\beta}}$, $b(t) = \frac{r-2\alpha}{t^{\alpha-\beta+1}}$, from conservation law (8) we get

$$\begin{aligned}
E &\equiv \frac{1}{2t_0^{\alpha-\beta}} \left\| \dot{W}(t_0) \right\|^2 + \frac{\alpha(\alpha+1-r)}{2t_0^{\alpha-\beta+2}} \|W(t_0)\|^2 + t_0^{\alpha+\beta} \left(f\left(\frac{W(t_0)}{t_0^\alpha} + X_\star \right) - f_\star \right) \\
&= \frac{1}{2t^{\alpha-\beta}} \left\| \dot{W} \right\|^2 + \frac{\alpha(\alpha+1-r)}{2t^{\alpha-\beta+2}} \|W\|^2 + t^{\alpha+\beta} \left(f\left(\frac{W}{t^\alpha} + X_\star \right) - f_\star \right) \\
&\quad + \int_{t_0}^t \frac{2r - 3\alpha - \beta}{s^{\alpha-\beta+1}} \left\| \dot{W} \right\|^2 ds + \int_{t_0}^t \frac{\alpha(\alpha+1-r)(\alpha-\beta+2)}{2s^{\alpha-\beta+3}} \|W\|^2 \, ds \\
&\quad - \int_{t_0}^t s^{\alpha+\beta-1} \left((\alpha+\beta) \left(f\left(\frac{W}{s^\alpha} + X_\star \right) - f_\star \right) - \alpha \left\langle \nabla f\left(\frac{W}{s^\alpha} + X_\star \right), \frac{W}{s^\alpha} \right\rangle \right) ds.
\end{aligned}$$

Rewriting in terms of X we have

$$\begin{aligned}
E &\equiv t_0^{\alpha+\beta}(f(X(t_0))) - f_\star) + \frac{1}{2} t_0^{\alpha+\beta-2} \left\| t_0 \dot{X}(t_0) + \alpha(X(t_0) - X_\star) \right\|^2 + \frac{1}{2} \alpha(\alpha+1-r) t_0^{\alpha+\beta-2} \|X(t_0) - X_\star\|^2 \\
&= t^{\alpha+\beta}(f(X) - f_\star) + \frac{1}{2} t^{\alpha+\beta-2} \left\| t\dot{X} + \alpha(X - X_\star) \right\|^2 + \frac{1}{2} \alpha(\alpha+1-r) t^{\alpha+\beta-2} \|X - X_\star\|^2 \\
&\quad + \int_{t_0}^t \frac{2r - 3\alpha - \beta}{2} s^{\alpha+\beta-3} \left\| s\dot{X} + \alpha(X - X_\star) \right\|^2 + \int_{t_0}^t \frac{\alpha(\alpha+1-r)(\alpha-\beta+2)}{2} s^{\alpha+\beta-3} \|X - X_\star\|^2 ds \\
&\quad + \int_{t_0}^t s^{\alpha+\beta-1} \Big((\alpha+\beta)(f_\star - f(X)) - \alpha \langle \nabla f(X), X_\star - X \rangle \Big) ds. \qquad (14)
\end{aligned}$$

To utilize the $\mathbf{H}_1(\gamma)$ hypothesis, it is natural to choose α, β such that $\frac{\alpha}{\alpha+\beta} = \frac{1}{\gamma}$. The choice $\alpha = \frac{2r}{\gamma+2}$, $\beta = \frac{2(\gamma-1)r}{\gamma+2}$ makes $\frac{\alpha}{\alpha+\beta} = \frac{1}{\gamma}$, and $2r - 3\alpha - \beta = 0$, and we get the conservation law used in Section 3.3.

$$\begin{aligned}
E &\equiv t^{\frac{2\gamma r}{\gamma+2}} (f(X) - f_\star) + \frac{1}{2} t^{\frac{2\gamma r}{\gamma+2} - 2} \left\| t\dot{X} + \frac{2r}{\gamma+2} (X - X_\star) \right\|^2 + \frac{r(2 - \gamma(r-1))}{(\gamma+2)^2} t^{\frac{2\gamma r}{\gamma+2} - 2} \|X - X_\star\|^2 \\
&\quad + \int_{t_0}^t \frac{2r(2r + 2 - \gamma(r-1))(2 - \gamma(r-1))}{(\gamma+2)^3} s^{\frac{2\gamma r}{\gamma+2} - 3} \|X - X_\star\|^2 ds \\
&\quad + \int_{t_0}^t s^{\frac{2\gamma r}{\gamma+2} - 1} \frac{2\gamma r}{\gamma+2} \left(f_\star - f(X) - \frac{1}{\gamma} \langle \nabla f(X), X_\star - X \rangle \right) ds.
\end{aligned}$$

C.3.1. LYAPUNOV FUNCTION FOR $r > 3$ IN (SU ET AL., 2014)

Plugging $\alpha = r-1$, $\beta = 3-r$, $t_0 = 0$ to (14), we have

$$\begin{aligned} E &\equiv \frac{(r-1)^2}{2}\left\|X_0 - X_\star\right\|^2 \\ &= t^2(f(X)-f_\star) + \frac{1}{2}\left\|t\dot{X} + (r-1)(X-X_\star)\right\|^2 \\ &\quad + \int_0^t s(r-1)\Big(f_\star - f(X) - \langle \nabla f(X), X_\star - X\rangle\Big)ds + \int_0^t s(r-3)(f(X)-f_\star)ds. \end{aligned}$$

Since all terms are nonnegative, we immediately get

$$f(X) - f_\star \le \frac{(r-1)^2}{2t^2}\left\|X_0 - X_\star\right\|^2.$$

In (Su et al., 2014), they also present

$$\int_0^\infty t(f(X(t)) - f^\star)dt \le \frac{(r-1)^2}{2(r-3)}\left\|X_0 - X^\star\right\|^2,$$

and this can also be obtained immediately from conservation law.

C.4. SC-AGM ODE

We proceed the argument similar to C.1. Start with the ODE (9)

$$0 = \ddot{X} + 2\sqrt{\mu}\dot{X} + \nabla f(X).$$

Now consider the coordinate change $W = e^{\beta t}(X - X_\star)$. Then we see

$$\begin{aligned} W &= e^{\beta t}(X - X_\star) \\ \dot{W} &= e^{\beta t}\left(\dot{X} + \beta(X - X_\star)\right) \\ \ddot{W} &= e^{\beta t}\left(\ddot{X} + 2\beta\dot{X} + \beta^2(X - X_\star)\right). \end{aligned}$$

From this, we can rewrite X, $\dot{X}$, $\ddot{X}$ in terms of W, $\dot{W}$, $\ddot{W}$,

$$\begin{aligned} X &= e^{-\beta t}W + X_\star \\ \dot{X} &= e^{-\beta t}\left(\dot{W} - \beta W\right) \\ \ddot{X} &= e^{-\beta t}\left(\ddot{W} - 2\beta\dot{W} + \beta^2 W\right). \end{aligned}$$

Plugging these to (9) we get ODE

$$0 = e^{-\beta t}\left(\ddot{W} + 2(\sqrt{\mu} - \beta)\dot{W} + \beta(\beta - 2\sqrt{\mu})W\right) + \nabla f\left(e^{-\beta t}W + X_\star\right).$$

Now by defining

$$U(W,t) = \frac{\beta(\beta - 2\sqrt{\mu})}{2}e^{-\beta t}\left\|W\right\|^2 + e^{\beta t}\left(f\left(e^{-\beta t}W + X_\star\right) - f_\star\right),$$

we can rewrite the ODE as

$$0 = e^{-\beta t}\ddot{W} + 2(\sqrt{\mu} - \beta)e^{-\beta t}\dot{W} + \nabla_W U(W,t).$$

Now plugging $a(t) = e^{-\beta t}$, $b(t) = 2(\sqrt{\mu} - \beta)e^{-\beta t}$, from conservation law (8) we get

$$\begin{aligned}
E &\equiv \frac{e^{-\beta t_0}}{2}\left\|\dot{W}(t_0)\right\|^2 + \frac{\beta(\beta - 2\sqrt{\mu})}{2}e^{-\beta t_0}\left\|W(t_0)\right\|^2 + e^{\beta t_0}\left(f(e^{-\beta t_0}W(t_0) + X_\star) - f_\star\right) \\
&= \frac{e^{-\beta t}}{2}\left\|\dot{W}\right\|^2 + \frac{\beta(\beta - 2\sqrt{\mu})}{2}e^{-\beta t}\left\|W\right\|^2 + e^{\beta t}\left(f\left(e^{-\beta t}W + X_\star\right) - f_\star\right) + \int_{t_0}^{t}\frac{4\sqrt{\mu} - 3\beta}{2}e^{-\beta s}\left\|\dot{W}\right\|^2 ds \\
&\quad - \int_{t_0}^{t}\left(\beta e^{\beta s}\left(f(e^{-\beta s}W + X_\star) - f_\star - \left\langle \nabla f(e^{-\beta s}W + X_\star), e^{-\beta s}W\right\rangle\right) - \frac{\beta^2(\beta - 2\sqrt{\mu})}{2}e^{-\beta s}\left\|W\right\|^2\right) ds.
\end{aligned}$$

Plugging $t_0 = 0$ and rewriting in terms of X, $\dot{X}$, $\ddot{X}$ we have

$$\begin{aligned}
E &\equiv f\left(X_0\right) - f_\star + \beta(\beta - \sqrt{\mu})\left\|X_0 - X_\star\right\|^2 \\
&= e^{\beta t}\left(f\left(X\right) - f_\star + \frac{1}{2}\left\|\dot{X} + \beta(X - X_\star)\right\|^2 + \frac{\beta(\beta - 2\sqrt{\mu})}{2}\left\|X - X_\star\right\|^2\right) \\
&\quad + \int_0^t \frac{4\sqrt{\mu} - 3\beta}{2}e^{\beta s}\left\|\dot{X} + \beta(X - X_\star)\right\|^2 ds \\
&\quad + \int_0^t \beta e^{\beta s}\left(f_\star - f(X) - \langle \nabla f(X), X_\star - X\rangle + \frac{\beta(\beta - 2\sqrt{\mu})}{2}\left\|X - X_\star\right\|^2\right) ds.
\end{aligned}$$

Now plugging $\beta = \sqrt{\mu}$ we have

$$\begin{aligned}
E &\equiv f\left(X_0\right) - f_\star \\
&= e^{\sqrt{\mu}t}\left(f\left(X\right) - f_\star + \frac{1}{2}\left\|\dot{X} + \sqrt{\mu}(X - X_\star)\right\|^2 - \frac{\mu}{2}\left\|X - X_\star\right\|^2\right) \\
&\quad + \int_0^t \frac{\sqrt{\mu}}{2}e^{\sqrt{\mu}s}\left\|\dot{X} + \sqrt{\mu}(X - X_\star)\right\|^2 ds \\
&\quad + \int_0^t \sqrt{\mu}e^{\sqrt{\mu}s}\left(f_\star - f(X) - \langle \nabla f(X), X_\star - X\rangle - \frac{\mu}{2}\left\|X - X_\star\right\|^2\right) ds.
\end{aligned}$$

Finally, from

$$\begin{aligned}
&\int_0^t \frac{\sqrt{\mu}}{2}e^{\sqrt{\mu}s}\left\|\dot{X} + \sqrt{\mu}(X - X_\star)\right\|^2 ds \\
&= \int_0^t \left(\frac{\sqrt{\mu}}{2}e^{\sqrt{\mu}s}\left\|\dot{X}\right\|^2 + \frac{\mu}{2}e^{\sqrt{\mu}s}\left(2\left\langle \dot{X}, X - X_\star\right\rangle + \sqrt{\mu}\left\|X - X_\star\right\|^2\right)\right) ds \\
&= \int_0^t \left(\frac{\sqrt{\mu}}{2}e^{\sqrt{\mu}s}\left\|\dot{X}\right\|^2 + \frac{\mu}{2}\frac{d}{ds}\left(e^{\sqrt{\mu}s}\left\|X - X_\star\right\|^2\right)\right) ds \\
&= \int_0^t \frac{\sqrt{\mu}}{2}e^{\sqrt{\mu}s}\left\|\dot{X}\right\|^2 ds + \frac{\mu}{2}\left[e^{\sqrt{\mu}s}\left\|X - X_\star\right\|^2\right]_0^t \\
&= \int_0^t \frac{\sqrt{\mu}}{2}e^{\sqrt{\mu}s}\left\|\dot{X}\right\|^2 ds + \frac{\mu}{2}\left(e^{\sqrt{\mu}t}\left\|X - X_\star\right\|^2 - \left\|X_0 - X_\star\right\|^2\right).
\end{aligned}$$

we conclude

$$\begin{aligned}
E &\equiv f\left(X_0\right) - f_\star \\
&= e^{\sqrt{\mu}t}\left(f\left(X\right) - f_\star + \frac{1}{2}\left\|\dot{X} + \sqrt{\mu}(X - X_\star)\right\|^2\right) - \frac{\mu}{2}\left\|X_0 - X_\star\right\|^2 \\
&\quad + \int_0^t \frac{\sqrt{\mu}}{2}e^{\sqrt{\mu}s}\left\|\dot{X}\right\|^2 ds + \int_0^t \sqrt{\mu}e^{\sqrt{\mu}s}\left(f_\star - f(X) - \langle \nabla f(X), X_\star - X\rangle - \frac{\mu}{2}\left\|X - X_\star\right\|^2\right) ds.
\end{aligned}$$

C.5. Gradient flow

Recall, gradient flow was written as

$$0 = \dot{X} + \nabla f(X).$$

Consider the dilated coordinate $W = t(X - X_\star)$. Then we see

$$\begin{aligned} W &= t(X - X_\star) \\ \dot{W} &= t\dot{X} + (X - X_\star). \end{aligned}$$

Then X, $\dot{X}$ can be rewritten as

$$\begin{aligned} X &= \frac{W}{t} + X_\star \\ \dot{X} &= \frac{\dot{W}}{t} - \frac{W}{t^2}. \end{aligned}$$

Plugging these to ODE, we have

$$0 = \frac{\dot{W}}{t} - \frac{W}{t^2} + \nabla f\left(\frac{W}{t} + X_\star\right).$$

Now by defining

$$U(W,t) = -\frac{1}{2t^2}\|W\|^2 + t\left(f\left(\frac{W}{t} + X_\star\right) - f_\star\right),$$

we can rewrite ODE as

$$0 = \frac{\dot{W}}{t} + \nabla_W U(W,t).$$

Now plugging $a(t) = 0$, $b(t) = \frac{1}{t}$, from conservation law (8)

$$\begin{aligned} E &\equiv \lim_{t_0 \to 0} U(W(t_0), t_0) \\ &= \int_0^t \frac{1}{s}\left\|\dot{W}\right\|^2 ds + U(W,t) - \int_0^t \frac{\partial}{\partial s} U(W,s) ds \\ &= \int_0^t \frac{1}{s}\left\|\dot{W}\right\|^2 ds - \frac{1}{2t^2}\|W\|^2 + t\left(f\left(\frac{W}{t} + X_\star\right) - f_\star\right) \\ &\quad - \int_0^t \left(\frac{1}{s^3}\|W\|^2 + \left(f\left(\frac{W}{s} + X_\star\right) - f_\star + s\left\langle \nabla f\left(\frac{W}{s} + X_\star\right), -\frac{W}{s^2}\right\rangle\right)\right) ds. \end{aligned}$$

Rewriting in terms of X, $\dot{X}$, we get the conservation law in Section 3.5

$$\begin{aligned} E &\equiv -\frac{1}{2}\|X_0 - X_\star\|^2 \\ &= t\left(f(X) - f_\star\right) - \frac{1}{2}\|X - X_\star\|^2 \\ &\quad + \int_0^t \left(\frac{1}{s}\left\|s\dot{X} + (X - X_\star)\right\|^2 - \frac{1}{s}\|X - X_\star\|^2\right) ds - \int_0^t \left(f(X) - f_\star - \langle \nabla f(X), X - X_\star\rangle\right) ds \\ &= t\left(f(X) - f_\star\right) - \frac{1}{2}\|X - X_\star\|^2 \\ &\quad + \int_0^t \left(s\left\|\dot{X}\right\|^2 + \frac{d}{ds}\|X - X_\star\|^2\right) ds + \int_0^t \left(f_\star - f(X) - \langle \nabla f(X), X_\star - X\rangle\right) ds \\ &= t\left(f(X) - f_\star\right) + \frac{1}{2}\|X - X_\star\|^2 - \|X_0 - X_\star\|^2 + \int_0^t s\left\|\dot{X}\right\|^2 ds + \int_0^t \left(f_\star - f(X) - \langle \nabla f(X), X_\star - X\rangle\right) ds. \end{aligned}$$

D. Omitted calculations of Section 4

D.1. Derivation of OGM-G ODE

OGM-G in (Kim & Fessler, 2021) was presented as

$$\begin{aligned} x_k^+ &= x_k - \frac{1}{L}\nabla f(x_k) \\ x_{k+1} &= x_k^+ + \frac{(\theta_{K-k}-1)(2\theta_{K-(k+1)}-1)}{\theta_{K-k}(2\theta_{K-k}-1)}(x_k^+ - x_{k-1}^+) + \frac{2\theta_{K-(k+1)}-1}{2\theta_{K-k}-1}(x_k^+ - x_k). \end{aligned}$$

Plugging $x_k^+ = x_k - \frac{1}{L}\nabla f(x_k)$ to the second line and using the fact $\theta_{K-k} = \frac{K-k}{2} + o(K)$ we have

$$\begin{aligned} x_{k+1} &= x_k - \frac{1}{L}\nabla f(x_k) + \frac{(K-k-2+o(K))^2}{(K-k+o(K))(K-k-1+o(K))}\left(x_k - x_{k-1} - \frac{1}{L}\left(\nabla f(x_k) - \nabla f(x_{k-1})\right)\right) \\ &\quad - \frac{K-k-2+o(K)}{K-k-1+o(K)}\frac{1}{L}\nabla f(x_k) \\ &= x_k + \left(1 - \frac{3(K-k)+o(K)}{(K-k)^2+o(K)K}\right)(x_k - x_{k-1}) - \left(2 - \frac{1}{K-k+o(K)}\right)\frac{1}{L}\nabla f(x_k) \\ &\quad - \frac{1}{L}\frac{(K-k)^2+o(K)K}{(K-k)^2+o(K)K}(\nabla f(x_k) - \nabla f(x_{k-1})). \end{aligned}$$

Similar to (Su et al., 2014), we use the identification $\frac{1}{L} = h^2$, $t = kh$ and $x_k = X(kh)$. Moreover for fixed $T > 0$, we use identification $T = Kh$. Adding $-2x_k + x_{k-1}$ and dividing h^2 both sides we have

$$\begin{aligned} \frac{(x_{k+1}-x_k)-(x_k-x_{k-1})}{h^2} &= -\frac{3(Kh-kh)+o(K)h}{(Kh-kh)^2+o(K)Kh^2}\frac{x_k - x_{k-1}}{h} - \left(2 - \frac{h}{Kh-kh+o(K)h}\right)\nabla f(x_k) \\ &\quad - \frac{(Kh-kh)^2+o(K)Kh^2}{(Kh-kh)^2+o(K)Kh^2}(\nabla f(x_k) - \nabla f(x_{k-1})) \\ &= -\frac{3}{T-t}\frac{X(t)-X(t-h)}{h} - 2\nabla f(X(t)) - (\nabla f(X(t)) - \nabla f(X(t-h)) + o(K)h. \end{aligned}$$

Finally taking limit $h \to 0$, we obtain the desired ODE

$$0 = \ddot{X}(t) - \frac{3}{t-T}\dot{X}(t) + 2\nabla f(X(t)).$$

D.1.1. OGM-G ODE COINCIDES WITH THE ODE MODEL OF OBL-G♭

The method OBL-G♭ (Park & Ryu, 2021)

$$\begin{aligned} x_k^+ &= x_k - \frac{1}{L}\nabla f(x_k) \\ z_{k+1} &= z_k - \frac{1}{L}\frac{K-k+1}{2}\nabla f(x_k) \\ x_{k+1} &= \frac{K-k-2}{K-k+2}x_k^+ + \frac{4}{K-k+2}z_{k+1}. \end{aligned}$$

is a variant of OGM-G. Interestingly, the ODE model of OBL-G♭ exactly coincides with OGM-G ODE.

Note this method is written in the form with auxiliary sequence z_k, we derive the ODE in a different way. We take the same identification $\frac{1}{L} = h^2$, $Kh = T$, $kh = t$, $x_k = X(kh)$, $z_k = Z(kh)$. Then we may regard the method as a system of first-order ODEs. From z_k update, by taking limit $h \to 0$ we have

$$\frac{z_{k+1}-z_k}{h} = -\frac{Kh-kh+h}{2}\nabla f(x_k) \quad \overset{h\to 0}{\Longrightarrow} \quad \dot{Z}(t) = -\frac{T-t}{2}\nabla f(X).$$

From x_k update, dividing both sides by h, subtracting x_k^+ both sides and by taking limit $h \to 0$ we have

$$\frac{x_{k+1}-x_k}{h} = \frac{4}{Kh-kh+2h}(z_{k+1}-x_k) - \frac{Kh-kh-2h}{Kh-kh+2h}\nabla f(x_k)h \quad \overset{h\to 0}{\Longrightarrow} \quad \dot{X}(t) = \frac{4}{T-t}(Z(t)-X(t)). \tag{15}$$

Thus we get system of first-order ODEs. Now to derive a second-order ODE, multiplying $T-t$ to (15) and differentiating, we have

$$(T-t)\ddot{X}(t) - \dot{X}(t) = 4\left(\dot{Z}(t)-\dot{X}(t)\right) = 4\left(-\frac{T-t}{2}\nabla f(X) - \dot{X}(t)\right).$$

Dividing $T-t$ and organizing the result, we conclude

$$0 = \ddot{X}(t) - \frac{3}{t-T}\dot{X}(t) + 2\nabla f(X).$$

D.2. Conservation law for OGM-G ODE

We proceed argument similar to C.4. Start with ODE presented in Section 4.2

$$0 = \ddot{X} + \frac{r}{t-T}\dot{X} + 2\nabla f(X). \tag{16}$$

Now consider the coordinate change $W = (T-t)^\alpha (X - X_c)$.

Then we see

$$\begin{aligned}
W(t) &= (T-t)^\alpha (X(t)-X_c)\\
\dot{W}(t) &= (T-t)^\alpha \dot{X}(t) - \alpha(T-t)^{\alpha-1}(X(t)-X_c)\\
\ddot{W}(t) &= (T-t)^\alpha \ddot{X}(t) - 2\alpha(T-t)^{\alpha-1}\dot{X}(t) + \alpha(\alpha-1)(T-t)^{\alpha-2}(X(t)-X_c).
\end{aligned}$$

Note the sign flips while differentiating $(T-t)^\alpha$.

From this, we can rewrite X, $\dot{X}$, $\ddot{X}$ in terms of W, $\dot{W}$, $\ddot{W}$,

$$\begin{aligned}
X(t) &= (T-t)^{-\alpha}W(t) + X_c\\
\dot{X}(t) &= (T-t)^{-\alpha}\dot{W}(t) + \alpha(T-t)^{-\alpha-1}W(t)\\
\ddot{X}(t) &= (T-t)^{-\alpha}\ddot{W}(t) + 2\alpha(T-t)^{-\alpha-1}\dot{W}(t) + \alpha(\alpha+1)(T-t)^{-\alpha-2}W(t).
\end{aligned}$$

Plugging these to (9) we get ODE

$$0 = \frac{1}{(T-t)^\alpha}\ddot{W} + \frac{2\alpha-r}{(T-t)^{\alpha+1}}\dot{W} + \frac{\alpha(\alpha+1-r)}{(T-t)^{\alpha+2}}W + 2\nabla f\left(\frac{W}{(T-t)^\alpha} + X_c\right).$$

Now by defining

$$U(W,t) = \frac{\alpha(\alpha+1-r)}{2(T-t)^{\alpha+2}}\|W\|^2 + 2(T-t)^\alpha\left(f\left(\frac{W}{(T-t)^\alpha}+X_c\right) - f(X_c)\right)$$

we can rewrite the ODE as

$$0 = \frac{1}{(T-t)^\alpha}\ddot{W} + \frac{2\alpha-r}{(T-t)^{\alpha+1}}\dot{W} + \nabla_W U(W,t).$$

Now plugging $a(t) = \frac{1}{(T-t)^\alpha}$, $b(t) = \frac{2\alpha-r}{(T-t)^{\alpha+1}}$, from conservation law (8) we get

$$\begin{aligned}
E &\equiv \frac{1}{2(T-t_0)^\alpha}\left\|\dot{W}(t_0)\right\|^2 + \frac{\alpha(\alpha+1-r)}{2(T-t_0)^{\alpha+2}}\|W(t_0)\|^2 + 2(T-t_0)^\alpha\left(f\left(\frac{W(t_0)}{(T-t_0)^\alpha}+X_c\right) - f(X_c)\right)\\
&= \frac{1}{2(T-t)^\alpha}\left\|\dot{W}\right\|^2 + \frac{\alpha(\alpha+1-r)}{2(T-t)^{\alpha+2}}\|W\|^2 + 2(T-t)^\alpha\left(f\left(\frac{W}{(T-t)^\alpha}+X_c\right) - f(X_c)\right)\\
&\quad + \int_{t_0}^t \frac{3\alpha-2r}{2(T-s)^{\alpha+3}}\left\|\dot{W}\right\|^2 ds - \int_{t_0}^t \frac{\alpha(\alpha+1-r)(\alpha+2)}{2(T-s)^{\alpha+3}}\|W\|^2\, ds\\
&\quad - \int_{t_0}^t \frac{2\alpha}{(T-s)^{\alpha+1}}\left(f(X_c) - f\left(\frac{W}{(T-s)^\alpha}+X_c\right) - \left\langle \nabla f\left(\frac{W}{(T-s)^\alpha}+X_c\right), \frac{W}{(T-s)^\alpha}\right\rangle\right) ds.
\end{aligned}$$

Plugging $t_0 = 0$ and rewriting in terms of X, $\dot{X}$, $\ddot{X}$ we have

$$
\begin{aligned}
E &= 2T^{\alpha}\left(f(X_0) - f(X_c)\right) + \left(\frac{\alpha^2}{2} + \frac{\alpha(\alpha+1-r)}{2}\right) T^{\alpha-2}\left\|X_0 - X_c\right\|^2 \qquad (17)\\
&= 2(T-t)^{\alpha}\left(f\left(X\right) - f(X_c)\right) + \frac{1}{2}(T-t)^{\alpha-2}\left\|(T-t)\dot{X} - \alpha(X - X_c)\right\|^2 + \frac{\alpha(\alpha+1-r)}{2}(T-t)^{\alpha-2}\left\|X - X_c\right\|^2\\
&\quad + \int_0^t \left(\frac{3\alpha - 2r}{2}(T-s)^{\alpha-3}\left\|(T-s)\dot{X} - \alpha(X - X_c)\right\|^2 - \frac{\alpha(\alpha+1-r)(\alpha+2)}{2}(T-s)^{\alpha-3}\left\|X - X_c\right\|^2\right) ds\\
&\quad + \int_0^t (-2\alpha)(T-s)^{\alpha-1}\left(f(X_c) - f(X) - \langle \nabla f(X), X_c - X\rangle\right) ds.
\end{aligned}
$$

Now plugging $\alpha = -2$ we get the energy in Section 4.2, moreover plugging $r = -3$ we get the energy for $r = -3$ in Section 4.

D.3. Regularity of OGM-G ODE at terminal time T

Since the argument for $r = -3$ is exactly same for general r, we prove the statement for the general $r < 0$. We will present our proofs in following order.

(i) $\sup_{t\in[0,T)} \left\|\dot{X}(t)\right\|$ is bounded.

(ii) $X(t)$ can be continuously extended to T.

(iii) $\lim_{t\to T^-} \dot{X}(t) = 0$.

(iv) $\lim_{t\to T^-} \frac{\dot{X}(t)}{t-T} = -\frac{2}{1+r}\nabla f(X(T))$.

(v) $\lim_{t\to T^-} \ddot{X}(t) = -\frac{2}{1+r}\nabla f(X(T))$.

(i), (ii) holds for $r \le 0$, (iii) holds for $r < 0$, and (iv), (v) holds for $r < 0$ with $r \neq -1$.

D.3.1. $\sup_{t\in[0,T)} \left\|\dot{X}(t)\right\|$ IS BOUNDED IF $r \le 0$

Considering conservation law (17) with $\alpha = 0$, $X_c = X_0$, we have

$$
E \equiv 0 = \frac{1}{2}\left\|\dot{X}(t)\right\|^2 + 2(f(X(t)) - f(X_0)) - \int_0^t \frac{r}{T-s}\left\|\dot{X}(s)\right\|^2 ds. \qquad (18)
$$

Collecting the terms except the integrand, define $\Psi\colon [0,T) \to \mathbb{R}$ as

$$
\Psi(t) = \frac{1}{2}\left\|\dot{X}(t)\right\|^2 + 2(f(X(t)) - f(X_0)).
$$

Observe for $r \le 0$

$$
\dot{\Psi}(t) = \frac{r}{T-t}\left\|\dot{X}(t)\right\|^2 \le 0,
$$

so $\Psi(t)$ is a nonincreasing function. Thus $\Psi(t) \le \Psi(0) = 0$, and from the fact $f_\star = \inf_{x\in\mathbb{R}^n} f(x) > -\infty$, we have

$$
\left\|\dot{X}(t)\right\|^2 = 2\Psi(t) + 4(f(X_0) - f(X(t))) \le 4(f(X_0) - f_\star).
$$

Therefore $\sup_{t\in[0,T)} \left\|\dot{X}(t)\right\| \le 2\sqrt{f(X_0) - f_\star}$, we get the desired result.

D.3.2. $X(t)$ CAN BE CONTINUOUSLY EXTENDED TO T

We first prove $X(t)$ is uniformly continuous. From the result of D.3.1, we see

$$\|X(t) - X(t+\delta)\| = \left\| \int_t^{t+\delta} \dot{X}(s) ds \right\| \leq \int_t^{t+\delta} \left\| \dot{X}(s) \right\| ds \leq \int_t^{t+\delta} 2\sqrt{f(X_0) - f_\star} ds = 2\delta\sqrt{f(X_0) - f_\star}.$$

Thus for X is $2\sqrt{f(X_0) - f_\star}$-Lipschitz function, we can conclude X is uniformly continuous.

Now from the fact of basic analysis, we know for $D \subset \mathbb{R}^n$, uniformly continuous function $g : D \to \mathbb{R}^n$ can be extended continuously to $\bar{D}$. Therefore $X \colon [0, T) \to \mathbb{R}^n$ can be extended to $\overline{[0, T)} = [0, T]$, we get the desired result.

D.3.3. $\lim_{t\to T^-} \left\| \dot{X}(t) \right\| = 0$

We first prove the limit $\lim_{t\to T^-} \left\| \dot{X}(t) \right\|$ exists. From Ψ defined in D.3.1 we have

$$\left\| \dot{X}(t) \right\| = \sqrt{2\Psi(t) + 4(f(X_0) - f(X(t)))},$$

so it is enough to show $\lim_{t\to T^-} \Psi(t)$ and $\lim_{t\to T^-} f(X(t))$ exists. From D.3.2 we know $\lim_{t\to T^-} X(t)$ exists, thus from continuity of f, we have $\lim_{t\to T^-} f(X(t))$ exists. It remains to show $\lim_{t\to T^-} \Psi(t)$ exists.

Recall Ψ is nonincreasing. Moreover, since $f_\star = \inf_{x\in\mathbb{R}^n} f(x) > -\infty$ we have

$$\Psi(t) = \frac{1}{2} \left\| \dot{X}(t) \right\|^2 + 2(f(X(t)) - f(X_0)) \geq 2(f_\star - f(X_0)),$$

so Ψ is bounded below. Thus Ψ is nonincreasing and bounded below, by completeness of real numbers, we conclude $\lim_{t\to T^-} \Psi(t)$ exists. Therefore $\lim_{t\to T^-} \left\| \dot{X}(t) \right\|$ exists.

Now we prove $\lim_{t\to T^-} \left\| \dot{X}(t) \right\| = 0$. Let $C = \lim_{t\to T^-} \left\| \dot{X}(t) \right\| \geq 0$. Assume for contradiction that $C > 0$. Then there is $\epsilon > 0$ such that $T - \epsilon < s < T$ implies $\left\| \dot{X}(s) \right\| > \frac{C}{2}$. Thus for $t > T - \epsilon$, if $r \leq 0$ we have

$$\int_0^t \frac{r}{T-s} \left\| \dot{X}(s) \right\|^2 ds = \int_0^{T-\epsilon} \frac{r}{T-s} \left\| \dot{X}(s) \right\|^2 ds + \int_{T-\epsilon}^t \frac{r}{T-s} \left\| \dot{X}(s) \right\|^2 ds \leq \int_{T-\epsilon}^t \frac{C^2}{4} \frac{r}{T-s} ds.$$

Since $\lim_{t\to T^-} \int_{T-\epsilon}^t \frac{C^2}{4} \frac{r}{(T-s)} ds = -\infty$ if $r < 0$, we conclude $\lim_{t\to T^-} \int_0^t \frac{r}{T-s} \left\| \dot{X}(s) \right\|^2 ds = -\infty$ from above inequality. By the way from (18) we know $\Psi(t) = \int_0^t \frac{r}{T-s} \left\| \dot{X}(s) \right\|^2 ds$, but we have just observed above that $\Psi(t)$ is bounded below. This is a contradiction, we conclude $\lim_{t\to T^-} \left\| \dot{X}(t) \right\| = 0$.

D.3.4. $\lim_{t\to T^-} \frac{\dot{X}(t)}{t-T} = -\frac{2}{r+1} \nabla f(X(T))$

The key observation of the proof is

$$\frac{d}{dt} \left((T-t)^r \dot{X}(t) \right) = -2(T-t)^r \nabla f(X(t)).$$

We can check above is true from the ODE $0 = \ddot{X} + \frac{r}{t-T} \dot{X} + 2\nabla f(X)$. With this observation, we can handle the separated terms $\ddot{X}$ and $\dot{X}$ as one term.

Integrating both sides from 0 to t, we get

$$(T-t)^r \dot{X}(t) = -\int_0^t 2(T-s)^r \nabla f(X(s)) ds.$$

Multiplying $(T-t)^{-(r+1)}$, we get

$$\frac{\dot{X}(t)}{T-t} = -(T-t)^{-(r+1)} \int_0^t 2(T-t)^r \nabla f(X(s)) ds. \tag{19}$$

From (Rockafellar, 1970, Corollary 25.5.1), the fact f is convex and differentiable implies continuity of ∇f. From D.3.2, we see $\lim_{t\to T^-} \nabla f(X(t))$ exists. Moreover from D.3.3, we see the numerator for left hand side reaches to zero as $t \to T^-$. Therefore we can apply L'Hôpital's rule (componentwisely), for $r \neq -1$ we conclude

$$\lim_{t\to T^-} \frac{\dot{X}(t)}{T-t} = -\lim_{t\to T^-} \frac{\int_0^t 2(T-t)^r \nabla f(X(s))\, ds}{(T-t)^{r+1}} = \frac{2}{r+1} \lim_{t\to T^-} \nabla f(X(t)) = \frac{2}{r+1} \nabla f(X(T)).$$

By flipping the sign of both sides, we get the desired result.

D.3.5. $\lim_{t\to T^-} \ddot{X}(t) = -\frac{2}{r+1} \nabla f(X(T))$

From ODE (16) we have

$$\ddot{X}(t) = \frac{r}{T-t} \dot{X}(t) - 2\nabla f(X(t)).$$

We know the limit $t \to T^-$ for right hand side exists by D.3.4. Therefore $\lim_{t\to T^-} \ddot{X}(t)$ exists, by L'Hôpital's rule we have

$$\lim_{t\to T^-} \ddot{X}(t) = \lim_{t\to T^-} \frac{\dot{X}(t)}{t-T} = -\frac{2}{r+1} \nabla f(X(T)).$$

D.4. Correspondence with discrete analysis of OGM-G

Lee et al. (2021) presented Lyapunov function proof for convergence analysis of OGM-G. They first rewrote OGM-G with auxiliary sequence z_k as follows

$$\begin{aligned} x_k^+ &= x_k - \frac{1}{L} \nabla f(x_k) \\ z_{k+1} &= z_k - \frac{\theta_{K-k}}{L} \nabla f(x_k) \qquad (20) \\ x_{k+1} &= \frac{\theta^4_{K-(k+2)}}{\theta^4_{K-(k+1)}} x_k^+ + \left(1 - \frac{\theta^4_{K-(k+2)}}{\theta^4_{K-(k+1)}}\right) z_{k+1}. \qquad (21) \end{aligned}$$

Then they presented the Lyapunov function as follows

$$\begin{aligned} U_k = \frac{1}{\theta^2_{K-k}} \left(\frac{1}{2L} \|f(x_K)\|^2 + \frac{1}{2L} \|f(x_k)\|^2 + f(x_k) - f(x_K) - \langle \nabla f(x_k), x_k - x_{k-1}^+ \rangle \right) \qquad (22) \\ + \frac{L}{\theta^4_{K-k}} \langle z_k - x_{k-1}^+, z_k - x_K^+ \rangle. \end{aligned}$$

We claim there is a correspondence between this function and the Lyapunov function we've presented in Theorem 4.2. We use same identification as did in D.1, $\frac{1}{L} = h^2$, $kh = t$, $Kh = T$, $x_k = X(kh)$, $z_k = Z(kh)$. Then we derive continuous counterpart of U_k by dividing $2h^2$ then ignoring $o(K)h$ and $O(h)$.

We first calculate the continuous counterpart of z_k. Rewrite the update equation (20) as

$$x_{k+1} - x_k^+ = \left(1 - \frac{\theta^4_{K-(k+2)}}{\theta^4_{K-(k+1)}}\right)(z_{k+1} - x_k^+). \tag{23}$$

Dividing left hand side with h we observe,

$$\frac{x_{k+1} - x_k^+}{h} = \frac{x_{k+1} - x_k + h^2 \nabla f(x_k)}{h} = \frac{x_{k+1} - x_k}{h} + O(h) = \dot{X}(t) + O(h).$$

Then from the fact $\theta_{K-k} = \frac{K-k}{2} + o(K)$, we observe

$$\begin{aligned}\frac{1}{h}\left(1 - \frac{\theta^4_{K-(k+2)}}{\theta^4_{K-(k+1)}}\right) &= \frac{1}{h}\left(1 - \frac{(K-k-2+o(K))^4}{(K-k-1+o(K))^4}\right)\\ &= \frac{1}{h}\left(\frac{(2(K-k)+o(K))(2(K-k)^2 - 6(K-k) + o(K)K)}{(K-k)^4 + o(K)K^3}\right)\\ &= \frac{(2(Kh-kh)+o(K)h)(2(Kh-kh)^2 - 6(Kh-kh)h + o(K)Kh^2)}{(Kh-kh)^4 + o(K)K^3h^4}\\ &= \frac{(2(T-t)+o(K)h)(2(T-t)^2 - 6(T-t)h + o(K)Th)}{(T-t)^4 + o(K)T^3h} = \frac{4}{T-t} + o(K)h.\end{aligned}$$

Dividing (23) by h, applying above observations, corresponding z_{k+1} with $Z(t+h) = Z(t) + O(h)$ we have

$$\dot{X}(t) + O(h) = \frac{x_{k+1} - x_k^+}{h} = \frac{1}{h}\left(1 - \frac{\theta^4_{K-(k+2)}}{\theta^4_{K-(k+1)}}\right)(z_{k+1} - x_k^+) = \frac{4}{T-t}\left(Z(t) - X(t)\right) + O(h) + o(K)h.$$

Organizing with respect to Z, we have

$$Z(t) = \frac{T-t}{4}\dot{X}(t) + X(t) + O(h) + o(K)h.$$

Now to conclude the desired result, we observe the followings. First, observe the terms with gradient are $O(h)$. For example, $\frac{1}{2L}\left\|\nabla f(x_K)\right\|^2 = \frac{h^2}{2}\left\|\nabla f(x_K)\right\|^2 = O(h)$. With this observation, we see $x^+_{k-1} = x_{k-1} - \frac{1}{L}\nabla f(x_{k-1})$ can be replaced with x_{k-1}. Second, observe $h\theta_{K-k} = \frac{T-t}{2} + o(K)h$. Third, we correspond x_{k-1} with $X(t-h) = X(t) + O(h)$.

Plugging these to (22), and dividing by $2h^2$, we get

$$\begin{aligned}&\frac{U_k}{2h^2} = \frac{1}{2(h\theta_{K-k})^2}\left(f(x_k) - f(x_K) + O(h)\right) + \frac{1}{2(h\theta_{K-k})^4}\left\langle z_k - x_k + O(h), z_k - x_K + O(h)\right\rangle\\ &= \frac{2}{(T-t+o(K)h)^2}\left(f(X(t)) - f(X(T))\right) + \frac{8}{(T-t+o(K)h)^4}\left\langle Z(t) - X(t), Z(t) - X(T)\right\rangle + O(h)\\ &= \frac{2}{(T-t)^2}\left(f(X(t)) - f(X(T))\right) + \frac{1}{2(T-t)^4}\left\langle (T-t)\dot{X}(t), (T-t)\dot{X}(t) + 4(X(t) - X(T))\right\rangle + O(h) + o(K)h\\ &= \frac{2}{(T-t)^2}\left(f(X(t)) - f(X(T))\right) + \frac{1}{2(T-t)^4}\left(\left\|(T-t)\dot{X}(t) + 2(X(t) - X(T))\right\|^2 - 4\left\|X(t) - X(T)\right\|^2\right)\\ &\quad + O(h) + o(K)h.\end{aligned}$$

Ignoring $O(h)$ and $o(K)h$, we see $\frac{U_k}{2h^2}$ corresponds to the Lyapunov function defined in Theorem 4.2.

D.5. Details for Theorem 4.3

Recall by plugging $\alpha = -2$, $X_c = X(T)$, $t_0 = 0$ to (17), we obtained the conservation law presented in 4.2.

$$\begin{aligned}E &\equiv \frac{2}{T^2}(f(X_0) - f(X(T))) + \frac{r+3}{T^4}\left\|X_0 - X(T)\right\|^2\\ &= \frac{2}{(T-t)^2}\left(f(X) - f(X(T))\right) + \frac{1}{2(T-t)^4}\left\|(T-t)\dot{X} + 2(X - X(T))\right\|^2 + \frac{r+1}{(T-t)^4}\left\|X - X(T)\right\|^2\\ &\quad + \int_0^t \frac{(-(r+3))}{(T-s)^5}\left\|(T-s)\dot{X} + 2(X - X(T))\right\|^2 ds\\ &\quad + \int_0^t \frac{4}{(T-s)^3}\left(f(X(T)) - f(X) - \langle\nabla f(X), X(T) - X\rangle\right) ds.\end{aligned}$$

By collecting first three terms, define the Lyapunov function as

$$\Phi(t) = \frac{2}{(T-t)^2}\left(f(X) - f(X(T))\right) + \frac{1}{2(T-t)^4}\left\|(T-t)\dot{X} + 2(X - X(T))\right\|^2 + \frac{r+1}{(T-t)^4}\left\|X - X(T)\right\|^2.$$

From conservation law we know $\dot{E} = 0$, so we have

$$\dot{\Phi}(t) = \frac{r+3}{(T-t)^5}\left\|(T-t)\dot{X} + 2(X - X(T))\right\|^2 - \frac{4}{(T-t)^3}\left(f(X(T)) - f(X) - \langle \nabla f(X), X - X(T)\rangle\right) \le 0.$$

Note the first term is nonpositive since $r \le -3$. Especially $\Phi(0) \ge \lim_{t\to T^-}\Phi(t)$.

Now we calculate $\lim_{t\to T^-}\Phi(t)$. From D.3 we know $\lim_{t\to T^-}\frac{\dot{X}(t)}{t-T} = -\frac{2}{r+1}\nabla f(X(T))$. By applying L'Hôpital's rule we have

$$\lim_{t\to T^-}\frac{f(X(t)) - f(X(T))}{(T-t)^2} = \lim_{t\to T^-}\frac{\left\langle \nabla f(X(t)), \dot{X}(t)\right\rangle}{-2(T-t)} = \left\langle \nabla f(X(T)), \lim_{t\to T^-}\frac{\dot{X}(t)}{2(t-T)}\right\rangle = -\frac{1}{r+1}\left\|\nabla f(X(T))\right\|^2$$

$$\lim_{t\to T^-}\frac{X(t) - X(T)}{(T-t)^2} = \lim_{t\to T^-}\frac{\dot{X}(t)}{-2(T-t)} = \frac{1}{2}\lim_{t\to T^-}\frac{\dot{X}(t)}{t-T} = -\frac{1}{r+1}\nabla f(X(T)).$$

Therefore we get

$$\begin{aligned}
\lim_{t\to T^-}\Phi(t) &= \lim_{t\to T^-}\left(\frac{2\big(f(X) - f(X(T))\big)}{(T-t)^2} + \frac{1}{2}\left\|-\frac{\dot{X}}{t-T} + 2\frac{X - X(T)}{(T-t)^2}\right\|^2 + (r+1)\left\|\frac{X - X(T)}{(T-t)^2}\right\|^2\right) \\
&= -\frac{2}{r+1}\left\|\nabla f(X(T))\right\|^2 + \frac{1}{2}\left\|\frac{2}{r+1}\nabla f(X(T)) - \frac{2}{r+1}\nabla f(X(T))\right\|^2 + \frac{1}{r+1}\left\|\nabla f(X(T))\right\|^2 \\
&= \frac{1}{-(r+1)}\left\|\nabla f(X(T))\right\|^2.
\end{aligned}$$

Finally applying above calculation we have

$$\begin{aligned}
\frac{1}{-(r+1)}\left\|\nabla f(X(T))\right\|^2 = \lim_{t\to T^-}\Phi(t) \le \Phi(0) &= \frac{2}{T^2}(f(X_0) - f(X(T))) + \frac{r+3}{T^4}\left\|X_0 - X(T)\right\|^2 \\
&\le \frac{2}{T^2}\left(f(X_0) - f(X(T))\right).
\end{aligned}$$

This proves Theorem 4.3.

E. Proof of Theorem 5.1

Recall, with $\theta_k = \frac{k}{2}$ the discretized method was

$$\begin{aligned}
x_k^+ &= x_k - \frac{s}{2}\nabla f(x_k) \\
z_{k+1} &= z_k - s\theta_k\nabla f(x_k) \\
x_{k+1} &= \frac{\theta_k^2}{\theta_{k+1}^2}x_k^+ + \left(1 - \frac{\theta_k^2}{\theta_{k+1}^2}\right)z_{k+1},
\end{aligned} \tag{12}$$

and with $c_k = \frac{\theta_{k+1}}{\theta_{k+1}^2 - \theta_k^2}$ the Lyapunov function was

$$\Phi_k = 2c_k\theta_k^2\left(f(x_k) - f_\star - \frac{s}{4}\left\|\nabla f(x_k)\right\|^2\right) + \frac{1}{s}\left\|z_{k+1} - X_\star\right\|^2$$

for $k = 0, 1, \ldots$. We first prove $\Phi_{k+1} \le \Phi_k$, then we will get the desired result from $\Phi_k \le \Phi_0$.

(i) $\Phi_{k+1} \le \Phi_k$
For convenience, name

$$A_k = c_k\theta_k^2 = \frac{\theta_{k+1}}{\theta_{k+1}^2 - \theta_k^2}\theta_k^2.$$

Observe since $c_k = \frac{2(k+1)}{(k+1)^2-k^2} = \frac{2(k+1)}{2k+1} \geq 1$, we have $A_k \geq \theta_k^2$. From this we have

$$\begin{aligned}\frac{1}{s}\left\|z_{k+1}-X_\star\right\|^2 - \frac{1}{s}\left\|z_{k+2}-X_\star\right\|^2 &= 2\theta_{k+1}\left\langle \nabla f(x_{k+1}), z_{k+1}-X_\star\right\rangle - s\theta_{k+1}^2\left\|\nabla f(x_{k+1})\right\|^2 \\ &\geq 2\theta_{k+1}\left\langle \nabla f(x_{k+1}), z_{k+1}-X_\star\right\rangle - sA_{k+1}\left\|\nabla f(x_{k+1})\right\|^2.\end{aligned}$$

Applying this fact we have

$$\begin{aligned}\Phi_k - \Phi_{k+1} &= 2A_k\left(f(x_k)-f_\star-\frac{s}{4}\left\|\nabla f(x_k)\right\|^2\right) - 2A_{k+1}\left(f(x_{k+1})-f_\star-\frac{s}{4}\left\|\nabla f(x_{k+1})\right\|^2\right) \\ &\quad + \frac{1}{s}\left\|z_{k+1}-X_\star\right\|^2 - \frac{1}{s}\left\|z_{k+2}-X_\star\right\|^2 \\ &\geq 2A_k\left(f(x_k)-f_\star-\frac{s}{4}\left\|\nabla f(x_k)\right\|^2\right) - 2A_{k+1}\left(f(x_{k+1})-f_\star-\frac{s}{4}\left\|\nabla f(x_{k+1})\right\|^2\right) \\ &\quad + 2\theta_{k+1}\left\langle \nabla f(x_{k+1}), z_{k+1}-X_\star\right\rangle - sA_{k+1}\left\|\nabla f(x_{k+1})\right\|^2 \\ &= 2A_k\left(f(x_k)-f_\star-\frac{s}{4}\left\|\nabla f(x_k)\right\|^2\right) - 2A_{k+1}\left(f(x_{k+1})-f_\star+\frac{s}{4}\left\|\nabla f(x_{k+1})\right\|^2\right) \\ &\quad + 2\theta_{k+1}\left\langle \nabla f(x_{k+1}), z_{k+1}-X_\star\right\rangle \\ &= 2A_k\left(f(x_k)-f_\star-\frac{s}{4}\left\|\nabla f(x_k)\right\|^2\right) - 2A_k\left(f(x_{k+1})-f_\star+\frac{s}{4}\left\|\nabla f(x_{k+1})\right\|^2\right) \\ &\quad + 2\underbrace{(A_k - A_{k+1} + \theta_{k+1})}_{=\frac{(k+1)^2}{8k^2+16k+6}\geq 0}\left(f(x_{k+1})-f_\star+\frac{s}{4}\left\|\nabla f(x_{k+1})\right\|^2\right) \\ &\quad - 2\theta_{k+1}\left(f(x_{k+1})-f_\star+\frac{s}{4}\left\|\nabla f(x_{k+1})\right\|^2\right) + 2\theta_{k+1}\left\langle \nabla f(x_{k+1}), z_{k+1}-X_\star\right\rangle \\ &\geq 2A_k\left(f(x_k)-f_\star-\frac{s}{4}\left\|\nabla f(x_k)\right\|^2\right) - 2A_k\left(f(x_{k+1})-f_\star+\frac{s}{4}\left\|\nabla f(x_{k+1})\right\|^2\right) \\ &\quad - 2\theta_{k+1}\left(f(x_{k+1})-f_\star+\frac{s}{4}\left\|\nabla f(x_{k+1})\right\|^2\right) + 2\theta_{k+1}\left\langle \nabla f(x_{k+1}), z_{k+1}-X_\star\right\rangle \\ &= 2A_k\left(f(x_k)-f(x_{k+1})-\frac{s}{4}\left\|\nabla f(x_k)\right\|^2-\frac{s}{4}\left\|\nabla f(x_{k+1})\right\|^2\right) \\ &\quad + 2\theta_{k+1}\left(f_\star - f(x_{k+1}) - \left\langle \nabla f(x_{k+1}), X_\star - x_{k+1}\right\rangle - \frac{s}{4}\left\|\nabla f(x_{k+1})\right\|^2\right) \\ &\quad + 2\theta_{k+1}\left\langle \nabla f(x_{k+1}), z_{k+1}-x_{k+1}\right\rangle \\ &\overset{(a)}{\geq} 2A_k\left(f(x_k)-f(x_{k+1})-\frac{s}{4}\left\|\nabla f(x_k)\right\|^2-\frac{s}{4}\left\|\nabla f(x_{k+1})\right\|^2\right) \\ &\quad + 2\theta_{k+1}\left\langle \nabla f(x_{k+1}), z_{k+1}-x_{k+1}\right\rangle \\ &= 2A_k\left(f(x_k)-f(x_{k+1})-\frac{s}{4}\left\|\nabla f(x_k)\right\|^2-\frac{s}{4}\left\|\nabla f(x_{k+1})\right\|^2\right) \\ &\quad + 2\theta_{k+1}\left\langle \nabla f(x_{k+1}), \frac{\theta_k^2}{\theta_{k+1}^2-\theta_k^2}\left(x_{k+1}-x_k^+\right)\right\rangle \\ &= 2A_k\left(f(x_k)-f(x_{k+1})-\frac{s}{4}\left\|\nabla f(x_k)\right\|^2-\frac{s}{4}\left\|\nabla f(x_{k+1})\right\|^2\right) \\ &\quad + 2A_k\left\langle \nabla f(x_{k+1}), x_{k+1}-x_k+\frac{s}{2}\nabla f(x_k)\right\rangle \\ &= 2A_k\left(f(x_k)-f(x_{k+1})+\left\langle \nabla f(x_{k+1}), x_{k+1}-x_k\right\rangle - \frac{s}{4}\left\|\nabla f(x_k)-\nabla f(x_{k+1})\right\|^2\right) \overset{(b)}{\geq} 0.\end{aligned}$$

The inequalities (a) and (b) come from the fact $s \in \left(0, \frac{2}{L}\right]$ and L-smoothness of f.

(ii) From $\Phi_k \leq \Phi_0$, we have $f(x_k^+) - f_\star \leq \frac{k+\frac{1}{2}}{k+1}\frac{2\|X_0-X_\star\|^2}{sk^2}$
From $\theta_0 = 0$ we have $A_0 = 0$, and so $z_1 = z_0 + s\theta_0\nabla f(X_0) = z_0 = X_0$. Therefore

$$\Phi_0 = 2A_0 + \frac{1}{s}\left\|z_1 - X_\star\right\|^2 = \frac{1}{s}\left\|X_0 - X_\star\right\|^2$$

Now since f is L-smooth, for $s \in \left(0, \frac{2}{L}\right]$, we have

$$f(x_k^+) \leq f(x_k) - \frac{1}{2L} \left\|\nabla f(x_k)\right\|^2 \leq f(x_k) - \frac{s}{4} \left\|\nabla f(x_k)\right\|^2,$$

and so

$$2A_k \left(f(x_k^+) - f_\star\right) \leq 2A_k \left(f(x_k) - f_\star - \frac{s}{4} \left\|\nabla f(x_k)\right\|^2\right) \leq \Phi_k \leq \Phi_0 = \frac{1}{s} \left\|X_0 - X_\star\right\|^2.$$

Therefore, we conclude

$$\begin{aligned} f(x_k^+) - f_\star &\leq \frac{\left\|X_0 - X_\star\right\|^2}{2sA_k} = \left(\frac{\theta_{k+1}}{\theta_{k+1}^2 - \theta_k^2}\theta_k^2\right)^{-1} \frac{\left\|X_0 - X_\star\right\|^2}{2s} \\ &= \left(\frac{2k+1}{2(k+1)} \times \frac{4}{k^2}\right) \frac{\left\|X_0 - X_\star\right\|^2}{2s} \\ &= \frac{k+\frac{1}{2}}{k+1} \frac{2\left\|X_0 - X_\star\right\|^2}{sk^2}. \end{aligned}$$

Since $\frac{k+\frac{1}{2}}{k+1} \leq 1$, this implies $f(x_k^+) - f_\star \leq \frac{2\|X_0 - X_\star\|^2}{sk^2}$ as well. This proves Theorem 5.1. □

F. Time-dependent Hamiltonian

For the sake of completeness, we show how the dynamics is described through a Hamiltonian perspective. With the Hamiltonian

$$\begin{aligned} H(W, P, t) &= \langle P, \dot{W} \rangle - L(W, P, t) \\ &= \frac{t}{2}\|P\|^2 + t^3(f(X(W, t)) - f_\star), \end{aligned}$$

the dynamics of the Euler–Lagrange equation can be equivalently specified with

$$\begin{aligned} \dot{P} &= -\nabla_W H(W, P, t) = -t\nabla f(X(W, t)) \\ \dot{W} &= \nabla_P H(W, P, t) = tP. \end{aligned}$$

However, our setup differs from the classical setup in that the Lagrangian and the Hamiltonian explicitly depend on time. One consequence of this difference is that the Hamiltonian is not conserved:

$$\begin{aligned} \frac{d}{dt} H(W, P, t) &= \left\langle \dot{W}, \nabla_W H(W, P, t) \right\rangle + \left\langle \dot{P}, \nabla_P H(W, P, t) \right\rangle + \frac{\partial}{\partial t} H(W, P, t) \\ &= \langle \nabla_P H(W, P, t), \nabla_W H(W, P, t) \rangle + \langle -\nabla_W H(W, P, t), \nabla_P H(W, P, t) \rangle + \frac{\partial}{\partial t} H(W, P, t) \\ &= \frac{\partial}{\partial t} H(W, P, t) \neq 0. \end{aligned}$$

Since H is not conserved, the classical theory of symplectic integrators is not immediately applicable.