Unifying Nesterov's Accelerated Gradient Methods for Convex and Strongly Convex Objective Functions

Jungbin Kim [1] **Insoon Yang** [1]

Abstract

Although Nesterov's accelerated gradient method (AGM) has been studied from various perspectives, it remains unclear why the most popular forms of AGMs must handle convex and strongly convex objective functions separately. To address this inconsistency, we propose a novel unified framework for Lagrangians, ordinary differential equation (ODE) models, and algorithms. As a special case, our new simple momentum algorithm, which we call the *unified AGM*, seamlessly bridges the gap between the two most popular forms of Nesterov's AGM and has a superior convergence guarantee compared to existing algorithms for non-strongly convex objective functions. This property is beneficial in practice when considering ill-conditioned μ-strongly convex objective functions (with small μ). Furthermore, we generalize this algorithm and the corresponding ODE model to the higher-order non-Euclidean setting. Last but not least, our unified framework is used to construct the *unified AGM-G ODE*, a novel ODE model for minimizing the gradient norm of strongly convex functions.

1. Introduction

We consider the optimization problem

$$\min_{x\in\mathcal{X}} f(x),$$

where $\mathcal{X} \subseteq \mathbb{R}^n$ is a convex set and $f : \mathcal{X} \to \mathbb{R}$ is a continuously differentiable function whose gradient is L-Lipschitz continuous. For the sake of simplicity, we assume $\mathcal{X} = \mathbb{R}^n$ and the objective function f has a minimizer x^*.

[1]Department of Electrical and Computer Engineering, ASRI, Seoul National University, Seoul, South Korea. Correspondence to: Insoon Yang <insoonyang@snu.ac.kr>.

Proceedings of the 40^{th} International Conference on Machine Learning, Honolulu, Hawaii, USA. PMLR 202, 2023.

Nesterov acceleration. Combining gradient descent with momentum, Nesterov (1983) proposed the accelerated gradient method (AGM). In particular, the following two specific schemes are the most popular versions of AGM: If f is convex, then the scheme

$$\begin{aligned} y_k &= x_k + \frac{2}{k+1}(z_k - x_k) \\ x_{k+1} &= y_k - s\nabla f(y_k) \\ z_{k+1} &= z_k - \frac{s(k+1)}{2}\nabla f(y_k) \end{aligned} \tag{AGM-C}$$

with $s \le 1/L$ achieves an $O\left(1/k^2\right)$ convergence rate (Tseng, 2008). If f is μ-strongly convex, then the scheme

$$\begin{aligned} y_k &= x_k + \frac{\sqrt{\mu s}}{1+\sqrt{\mu s}}(z_k - x_k) \\ x_{k+1} &= y_k - s\nabla f(y_k) \\ z_{k+1} &= z_k + \sqrt{\mu s}\left(y_k - z_k - \frac{1}{\mu}\nabla f(y_k)\right) \end{aligned} \tag{AGM-SC}$$

with $s \le 1/L$ exhibits an $O((1-\sqrt{\mu s})^k)$ convergence rate (Nesterov, 2018). Here, we observe that

AGM-SC does not recover AGM-C as $\mu \to 0$,

which indicates the inconsistency between AGM-SC and AGM-C. Furthermore, when μ is very small, the convergence rate of AGM-SC is slower than AGM-C in the early stage because $(1-\sqrt{\mu s})^k$ tends to zero very slowly. This is unexpected, as AGM-SC exploits the strong convexity of f, while AGM-C only relies on the convexity of f.

Recent studies on Nesterov acceleration, particularly in the development of ODE models and Lagrangian formulations, have primarily focused on the algorithms AGM-C and AGM-SC due to their simplicity. As a result, the inconsistency between these two algorithms is inherited in the corresponding ODEs and Lagrangians (and this is further discussed in Appendix A.1). The main goal of this paper is to address such inconsistencies in the discrete-time algorithms, continuous-time dynamics and Lagrangians.

Practical perspective. We now explain the importance of designing efficient algorithms for minimizing μ-strongly

convex functions which are ill-conditioned (i.e., having a small strong convexity parameter μ). Many optimization problems in machine learning can be formulated as

$$\min_{x\in\mathbb{R}^n} \; f(x) = \frac{1}{m}\left(\sum_{i=1}^{m} f_i(x) + \lambda R(x)\right). \tag{1}$$

Consider the problem (1) with L-smooth convex loss functions f_i and ℓ_2-regularization term $R(x) = \|x\|^2$. Then, f is L-smooth and $\frac{2\lambda}{m}$-strongly convex. Since the strong convexity parameter μ decreases as the sample size m increases or the regularization parameter λ decreases, improving the convergence rate for ill-conditioned strongly convex objective functions is significant, as emphasized in (Bubeck et al., 2015, Section 3.6).

1.1. Contributions

In this paper, we propose a novel Lagrangian, continuous-time models and discrete-time algorithms that handle convex and strongly convex objective functions in a unified way. The main contributions of this work can be summarized as follows:

- We propose a novel Lagrangian, called the *unified Bregman Lagrangian* (4), that handles convex and strongly convex cases simultaneously. By solving the Euler–Lagrange equation, we obtain a family of momentum dynamics for minimizing both convex and strongly convex objective functions.
- As a special case of the unified Bregman Lagrangian flow, we derive an ODE model that minimizes both convex and strongly convex functions, called the *unified AGM ODE* (8). As a rate-matching discretization of this ODE model, we devise a novel algorithm, called the *unified AGM* (Algorithm 1). The proposed algorithm always has a better convergence guarantee than AGM-C and reduces to AGM-C when $\mu = 0$.
- We extend the unified AGM ODE to the higher-order non-Euclidean setting, resulting in a novel continuous-time model called the *unified ATM ODE* (12). As a rate-matching discretization of this ODE model, we propose a novel higher-order method for minimizing both convex and uniformly convex functions, called the *unified ATM* (Algorithm 2).
- We strengthen the connection between our proposed unified dynamics/methods and existing ones by analyzing two limit cases: As $\mu \to 0$, our unified methods directly reduce to the ones for the non-strongly convex case. When $\mu > 0$, our unified methods recover the *time-invariant* methods for the strongly convex case as the *asymptotic limits*.

The following byproducts and observations are not directly related to our main goal but may be of independent interest:

- We develop a novel tool, called the *differential kernel*, which allows us to derive the limiting ODEs of fixed-step first-order methods (2). Our result recovers the limiting ODEs of the *three-sequence scheme* and *two-sequence scheme*, which are commonly found in the literature.
- Taking inspiration from the *anti-transpose relationship* between OGM ODE and OGM-G ODE, we propose the *unified AGM-G ODE* (17), a novel ODE model for minimizing the gradient norm of strongly convex functions.

1.2. Related Work

Nesterov's accelerated gradient methods are first proposed in (1983; 2018). Su et al. (2014; 2016) derived the limiting ODE of AGM-C, which has further been generalized and investigated in (Krichene et al., 2015; Attouch et al., 2018). Wibisono et al. (2016) developed the first Bregman Lagrangian, which systematically generates a family of continuous-time flows including the limiting ODE of AGM-C and its higher-order extensions. In the strongly convex case, Wilson et al. (2021) developed the second Bregman Lagrangian, which generates a family of continuous-time flows including the limiting ODE of AGM-SC. However, as discussed in Appendix A.1.2, their work is not consistent with (Wibisono et al., 2016). Based on the Lagrangian and Hamiltonian formulations, Betancourt et al. (2018); França et al. (2021); Muehlebach & Jordan (2021) studied a symplectic integrator to achieve acceleration in discrete-time settings. Shi et al. (2021) derived high-resolution ODEs for AGM-C and AGM-SC, from which Shi et al. (2019); Zhang et al. (2021) obtained accelerated methods by applying Euler methods. Diakonikolas & Orecchia (2019) proposed the approximate duality gap technique to understand AGM in both continuous-time and discrete-time settings. Notably, the *constant step scheme I* presented in (Nesterov, 2018, Equation 2.2.19) and the *NAG flow* presented in (Luo & Chen, 2021) are closely related to our work, as they handle the convex case and the strongly convex case simultaneously. In Appendices A.3 and D.8, we show that the algorithm and ODE model can be recovered from our unified framework.

Accelerated methods with higher-order tensor update step have been studied in (Nesterov, 2008; Baes, 2009; Wibisono et al., 2016; Gasnikov et al., 2019). Accelerated methods for reducing the gradient norm of objective functions have also been explored in several works (Nesterov, 2012; Ito & Fukuda, 2021). In particular, Kim & Fessler (2021) proposed the OGM-G, an optimal method for minimizing the

gradient norm of convex objective functions, which is further analyzed using Lyapunov arguments (Diakonikolas & Wang, 2022; Lee et al., 2021), and ODE models (Suh et al., 2022).

2. Preliminaries

In this section, we provide basic concepts and some novel tools that will be used throughout the paper.

2.1. Higher-Order Hyperbolic Functions

In this subsection, we introduce a family of *higher-order hyperbolic functions*, which are parametrized by the order $p = 2, 3, \ldots$. The definitions of these functions with $p \geq 3$ are developed for the purpose of designing the unified ATM ODE presented in Section 5.1. We define the $\sinh_p$ function as the solution to the problem

$$\sinh_p'(t) = \cosh_p(t) := (1 + \sinh_p^p(t))^{1/p}, \ \sinh_p(0) = 0.$$

We define the $\tanh_p$, $\coth_p$, sech_p, and csch_p functions as

$$\tanh_p(t) := \frac{\sinh_p(t)}{\cosh_p(t)}, \quad \coth_p(t) := \frac{\cosh_p(t)}{\sinh_p(t)},$$
$$\operatorname{sech}_p(t) := \frac{1}{\cosh_p(t)}, \quad \operatorname{csch}_p(t) := \frac{1}{\sinh_p(t)}.$$

When $p = 2$, these functions reduce to the standard hyperbolic functions, in which case we omit the subscript p. Following ten Thije Boonkkamp et al. (2012), we define the sinhc_p, tanhc_p, cothc_p, and cschc_p functions as

$$\operatorname{sinhc}_p(t) := \frac{\sinh_p(t)}{t}, \quad \operatorname{cschc}_p(t) := \frac{t}{\sinh_p(t)},$$
$$\operatorname{tanhc}_p(t) := \frac{\operatorname{sinhc}_p(t)}{\cosh_p(t)}, \quad \operatorname{cothc}_p(t) := \frac{\cosh_p(t)}{\operatorname{sinhc}_p(t)}.$$

The graphs of these functions are shown in Figures 4 and 5.

2.2. Differential Kernel

In this subsection, we propose a novel tool, called the *differential kernel*, for deriving limiting ODEs of general first-order methods. For brevity, we only present the key results here and refer the readers to Appendix B.2.3 for detailed calculations. Most of the first-order momentum algorithms can be formulated as the following *fixed-step first-order scheme* (see Drori & Teboulle, 2014):

$$y_{i+1} = y_i - s\sum_{j=0}^{i} h_{ij}\nabla f\left(y_j\right). \tag{2}$$

To derive the limiting ODE of (2), we introduce the *ansatzes* $y_k \approx X(k\sqrt{s})$ and $h_{ij} \approx H(i\sqrt{s}, j\sqrt{s})$ for some smooth curve $X(t)$ and some smooth function $H(t,\tau)$. Then, taking the limit $s \to 0$ in (2) yields

$$\dot{X}(t) = -\int_0^t H(t,\tau)\nabla f(X(\tau))\, d\tau, \tag{3}$$

where $H(t,\tau) = \lim_{s\to 0} h_{t/\sqrt{s},\tau/\sqrt{s}}$. We call $H(t,\tau)$ the *differential kernel* (or *H-kernel*) for the integro-differential equation (3). Note that the form of (3) clearly shows the momentum effect as it shows that the gradient $\nabla f(X(\tau))$ at time τ influences the velocity $\dot{X}(t)$ at all future times $t > \tau$.

Equivalence with second-order ODE. The equation (3) may seem different from the typical ODE models found in the literature, such as AGM-C ODE presented in (Su et al., 2016). However, it is possible to convert the equation (3) into a second-order ODE under certain conditions. If there exists a function $b(t)$ such that

$$\frac{\partial H(t,\tau)}{\partial t} = -b(t)H(t,\tau),$$

then the equation (3) can be rewritten as the following second-order ODE (see Appendix B.2.3):

$$\ddot{X}(t) + b(t)\dot{X} + H(t,t)\nabla f(X(t)) = 0.$$

As concrete examples, we can readily verify that (3) with the differential kernel $H^{\mathrm{C}}(t,\tau) = \tau^3/t^3$ is equivalent to AGM-C ODE and that (3) with the differential kernel $H^{\mathrm{SC}}(t,\tau) = e^{2\sqrt{\mu}(\tau-t)}$ is equivalent to AGM-SC ODE.

3. Unified Lagrangian Formulation

In this section, we propose a novel Lagrangian framework that can handle both non-strongly convex ($\mu = 0$) and strongly convex ($\mu > 0$) cases, unlike the existing frameworks in (Wibisono et al., 2016; Wilson et al., 2021), which are limited to either $\mu = 0$ or $\mu > 0$. In Section 3.2, we discuss that our novel framework is closely related to the aforementioned existing frameworks.

3.1. Unified Bregman Lagrangian

For a differentiable, convex, and essentially smooth function $h : \mathbb{R}^n \to \mathbb{R}$ and continuously differentiable functions $\alpha, \beta, \gamma : [0,\infty) \to \mathbb{R}$ satisfying the *ideal scaling conditions* $\dot{\gamma}(t) = e^{\alpha(t)}$ and $\dot{\beta}(t) \leq e^{\alpha(t)}$, we define the *unified Bregman Lagrangian* as[1]

$$\mathcal{L}(X,\dot{X},t) = e^{\alpha(t)+\gamma(t)}((1+\mu e^{\beta(t)}) \times D_h(X + e^{-\alpha(t)}\dot{X}, X) - e^{\beta(t)}f(X)), \tag{4}$$

[1]This Lagrangian recovers the first Bregman Lagrangian (22) when $\mu = 0$ unlike the second Bregman Lagrangian (26).

where D_h is the *Bregman divergence* of h defined as $D_h(y,x) = h(y) - h(x) - \langle \nabla h(x), y - x \rangle$. In Appendix C.1, we show that the Euler–Lagrange equation $\frac{d}{dt}\frac{\partial \mathcal{L}}{\partial \dot{X}} = \frac{\partial \mathcal{L}}{\partial X}$ for the unified Bregman Lagrangian (4) reduces to the following system of ODEs, which we call the *unified Bregman Lagrangian flow*:

$$\begin{aligned} \dot{X} &= e^{\alpha}(Z - X) \\ \frac{d}{dt}\nabla h(Z) &= \frac{\mu \dot{\beta} e^{\beta}}{1 + \mu e^{\beta}}(\nabla h(X) - \nabla h(Z)) \\ &\quad - \frac{e^{\alpha+\beta}}{1 + \mu e^{\beta}} \nabla f(X). \end{aligned} \tag{5}$$

The convergence rate of this ODE model is addressed in the following theorem.

Theorem 3.1. *Let f be a μ-uniformly (possibly with $\mu = 0$) convex function with respect to h.[2] Then, any solution to the system of ODEs (5) satisfies*

$$\begin{aligned} f(X(t)) - f(x^*) &\le e^{-\beta(t)}\big((1 + \mu e^{\beta(0)}) \\ &\quad \times D_h(x^*, Z(0)) + e^{\beta(0)}(f(X(0)) - f(x^*))\big). \end{aligned} \tag{6}$$

The proof of Theorem 3.1 can be found in Appendix C.3.

3.2. Limit Cases of Unified Bregman Lagrangian Flow

In this subsection, we analyze two limit cases of the unified Bregman Lagrangian flow.[3] First, when $\mu = 0$, the flow (5) reduces to the *first Bregman Lagrangian flow* (25), a known ODE model for the non-strongly convex case. Second, when $\mu > 0$ and the limits $\alpha(\infty)$ and $\dot{\beta}(\infty) > 0$ exist, we argue that the flow (5) is closely related to the following system of ODEs:

$$\begin{aligned} \dot{X} &= e^{\alpha(\infty)}(Z - X) \\ \frac{d}{dt}\nabla h(Z) &= \dot{\beta}(\infty)\left(\nabla h(X) - \nabla h(Z)\right) - \frac{e^{\alpha(\infty)}}{\mu}\nabla f(X), \end{aligned} \tag{7}$$

which is the *second Bregman Lagrangian flow* (27), a known ODE model for the strongly convex case, with the parameters $\alpha_{\mathrm{2nd}}(t) :\equiv \alpha(\infty)$ and $\beta_{\mathrm{2nd}}(t) := \dot{\beta}(\infty)t$. Consider the dynamics (5) and (7) as systems whose inputs are the initial point $x_0 = X(t_0) = Z(t_0)$ at the initial time t_0 and outputs are $X(t_0 + T)$ at the final time $t_0 + T$. Then, for fixed x_0, it can be shown that the output of the system (5) converges to the output of the system (7) as $t_0 \to \infty$ (with a formal statement and its proof found in Appendix C.4). In light of this, the system (7) is regarded as the *asymptotic limit* of the system (5). Furthermore, in Appendix C.5, we show that the convergence rate of (7) can be recovered from the convergence analysis of (5) via a limiting argument. The system (7) is *time-invariant*.

[2] We say that f is μ-uniformly convex with respect to h if the inequality $\mu D_h(x,y) \le D_f(x,y)$ holds for all $x, y \in \mathbb{R}^n$.

[3] This analysis is inspired by the two limit cases of ITEM (a discrete-time algorithm) in (Taylor & Drori, 2022, Section 2.2). However, our analysis is more rigorous and detailed.

4. Unified ODE Model and Algorithm

In this section, using the unified Lagrangian framework, we propose a novel algorithm that unifies AGM-C and AGM-SC, for minimizing both convex and strongly convex functions. Throughout this section, we assume that the objective function f is L-smooth and μ-strongly (possibly with $\mu = 0$) convex.

4.1. Proposed Dynamics: Unified AGM ODE

We consider the unified Bregman Lagrangian flow (5) with $h(x) = \frac{1}{2}\|x\|^2$, $\alpha(t) = \log\left(\frac{2}{t}\operatorname{cothc}(\frac{\sqrt{\mu}}{2}t)\right)$, and $\beta(t) = \log\left(\frac{t^2}{4}\operatorname{sinhc}^2(\frac{\sqrt{\mu}}{2}t)\right)$,[4] which can be equivalently written as the following second-order ODE (see Appendix D.2):

$$\begin{aligned} \ddot{X} + \Bigg(\frac{\sqrt{\mu}}{2}\tanh\left(\frac{\sqrt{\mu}}{2}t\right) \\ + \frac{3}{t}\operatorname{cothc}\left(\frac{\sqrt{\mu}}{2}t\right)\Bigg)\dot{X} + \nabla f(X) = 0, \end{aligned} \tag{8}$$

which we call the *unified AGM ODE*. This ODE has a unique solution as shown in Appendix D.3. The following theorem describes the convergence rate of the dynamics (5).

Theorem 4.1. *The solution to the unified AGM ODE (8) with the initial conditions $X(0) = x_0$ and $\dot{X}(0) = 0$ satisfies*

$$\begin{aligned} f(X(t)) - f(x^*) &\le \frac{2}{t^2}\operatorname{cschc}^2\left(\frac{\sqrt{\mu}}{2}t\right)\|x_0 - x^*\|^2 \\ &= O\big(\min\{1/t^2, e^{-\sqrt{\mu}t}\}\big). \end{aligned} \tag{9}$$

The proof of Theorem 4.1 can be found in Appendix D.4.

4.2. Proposed Algorithm: Unified AGM

We present the *unified AGM*, which is a rate-matching discretization of the unified AGM ODE (8), in Algorithm 1. This algorithm achieves an accelerated convergence rate as shown in the following theorem.

Theorem 4.2. *The iterates of the unified AGM (Algorithm 1) with $s \le 1/L$ satisfy*

$$\begin{aligned} f(x_k) - f(x^*) &\le \frac{2}{\iota^2 s k^2}\operatorname{cschc}^2\left(\frac{\iota\sqrt{\mu s}}{2}k\right)\|x_0 - x^*\|^2 \\ &= O\big(\min\{1/k^2, (1 - \sqrt{\mu s})^k\}\big). \end{aligned} \tag{10}$$

[4] These functions are chosen constructively as described in Appendix D.1.

[5] Note that the parameter ι is continuous in μ, as $\lim_{\mu \to 0}\iota = 1$.

Algorithm 1 Unified AGM

Input: Initial point $x_0 \in \mathbb{R}^n$, stepsize s
Initialize $z_0 = x_0$; $q = \mu s$
if $\mu = 0$ **then** $\iota = 1$ **else** $\iota = -\frac{\log(1-\sqrt{\mu s})}{\sqrt{\mu s}}$
for $k = 0, 1, 2, \dots$ **do**
$y_k = x_k + \frac{1}{1-q}\left(\frac{2}{\iota(k+1)}\operatorname{cothc}\left(\frac{k+1}{2}\iota\sqrt{q}\right) - q\right)(z_k - x_k)$
$x_{k+1} = y_k - s\nabla f(y_k)$
$z_{k+1} = z_k + \frac{\iota s(k+1)}{2}\operatorname{tanhc}\left(\frac{k+1}{2}\iota\sqrt{q}\right) \times(\mu y_k - \mu z_k - \nabla f(y_k))$
end for

The proof of Theroem 4.2 can be found in Appendix D.5. The unified AGM exhibits the best of both polynomial $O(1/k^2)$ and exponential $O((1-\sqrt{\mu s})^k)$ convergence rates, while each of AGM-C and AGM-SC achieves only one of these rates.

Unified AGM converges to unified AGM ODE. In Appendix D.6, we show that the iterates of the unified AGM converge to the solution to the unified AGM ODE under the identifications $t \leftrightarrow \mathbf{t}_k$, $X(\mathbf{t}_k) \leftrightarrow x_k$, and $Z(\mathbf{t}_k) \leftrightarrow z_k$, where $\mathbf{t}_k := \iota\sqrt{s}k$ and $Z(t) = X(t) + \frac{t}{2}\dot{X}(t)$. Because the convergence rates (9) and (10) are equivalent under these identifications, the unified AGM is a rate-matching discretization of the unified AGM ODE.

4.3. Limit Cases of Unified AGM ODE and Unified AGM

We now examine the two limit cases of the proposed ODE and algorithm. When $\mu = 0$, the unified AGM ODE (8) is simplified to

$$\ddot{X} + \frac{3}{t}\dot{X} + \nabla f(X) = 0, \qquad \text{(AGM-C ODE)}$$

which is the limiting ODE of AGM-C (see Su et al., 2016). When $\mu > 0$, the asymptotic limit of the unified AGM ODE is the system (7) with $\alpha(\infty) = \log\sqrt{\mu}$ and $\dot{\beta}(\infty) = \sqrt{\mu}$, which can be equivalently written as

$$\ddot{X} + 2\sqrt{\mu}\dot{X} + \nabla f(X) = 0. \qquad \text{(AGM-SC ODE)}$$

Note that this is the limiting ODE of AGM-SC (see Wilson et al., 2021). The coefficient of $\dot{X}$ in the unified AGM ODE converges pointwise to $\frac{3}{t}$ as $\mu \to 0$ and converges to $2\sqrt{\mu}$ as $t \to \infty$ (see Figure 1). This observation is aligned with the limiting arguments above.

When $\mu = 0$, the unified AGM simply reduces to AGM-C. When $\mu > 0$, viewing the unified AGM and AGM-SC as systems whose inputs are the initial point $x_{k_0} = z_{k_0}$ at the initial iteration k_0 and the outputs are x_{k_0+K} at the final iteration $k_0 + K$, we can show that AGM-SC is the *asymptotic limit* of the unified AGM. This means that the output

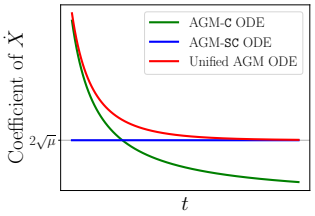


Figure 1. Plots for the coefficient of $\dot{X}$.

of the unified AGM converges to the output of AGM-SC as $k_0 \to \infty$ (with a formal statement and its proof found in Appendix D.7).

Unified AGM is always better than AGM-C. The convergence rate (10) of the unified AGM improves as μ increases and exactly recovers the convergence rate of AGM-C when $\mu = 0$. Thus, the unified AGM always has a better convergence guarantee than AGM-C. In contrast, the convergence guarantee of AGM-SC is no better than that of AGM-C when μ is small.

Differential kernel of the unified AGM ODE. The differential kernel of the unified AGM ODE (8) can be computed as (see Appendix D.2)

$$H^{\mathrm{U}}(t,\tau) = \frac{\tau^3 \operatorname{sinhc}^3\left(\frac{\sqrt{\mu}}{2}\tau\right)\cosh\left(\frac{\sqrt{\mu}}{2}\tau\right)}{t^3 \operatorname{sinhc}^3\left(\frac{\sqrt{\mu}}{2}t\right)\cosh\left(\frac{\sqrt{\mu}}{2}t\right)}. \qquad (11)$$

When $\mu = 0$, $H^{\mathrm{U}}(t,\tau)$ reduces to the differential kernel of AGM-C ODE, $H^{\mathrm{C}}(t,\tau) = \tau^3/t^3$. When $\mu > 0$, $H^{\mathrm{U}}(t,\tau)$ is asymptotically equivalent to $H^{\mathrm{SC}}(t,\tau) = e^{2\sqrt{\mu}(\tau-t)}$, the differential kernel of AGM-SC ODE, that is, $H^{\mathrm{U}}(t,\tau)/H^{\mathrm{SC}}(t,\tau) \to 1$ as $t, \tau \to \infty$. This is consistent with the fact that AGM-SC ODE is the asymptotic limit of the unified AGM ODE.

5. Unified Higher-Order Method

Using the *first Bregman Lagrangian* (an existing Lagrangian framework for the non-strongly convex case), Wibisono et al. (2016) proposed the *accelerated tensor method* (ATM-C) and its limiting ODE (ATM-C ODE) for convex objective functions, achieving $O(1/k^p)$ and $O(1/t^p)$ convergence rates, respectively. The authors also attempted to extend their results to the uniformly convex case within the first Bregman Lagrangian framework, where the goal is to achieve an exponential convergence rate. However, only an ODE model was proposed, and its rate-matching discretization was not identified. Instead, they showed that ATM-C with a restart scheme achieves an accelerated exponential convergence rate. Despite this progress, as they mentioned, making the connection between discrete-time algorithms

and continuous-time flows in the uniformly convex case remains an open problem.

In this section, our unified Lagrangian framework is used to seamlessly extend ATM-C and ATM-C ODE to the uniformly convex case. Our novel ODE model and algorithm achieve exponential convergence rates without relying on a restart scheme. Throughout this section, we assume that the distance-generating function h is 1-uniformly convex of order p,[6] the objective function f is p-times continuously differentiable and L-smooth of order $p-1$,[7] and that f is μ-uniformly convex with respect to h (possibly with $\mu = 0$).[8] These assumptions are standard in the literature about higher-order optimization in the mirror descent setup (see Wibisono et al., 2016; Wilson et al., 2021) and recover the setting in Section 4 when $p = 2$ and $h(x) = \|x\|^2$.

5.1. Proposed Dynamics: Unified ATM ODE

We consider the unified Bregman Lagrangian flow (5) with the parameters $\alpha(t) = \log(\frac{p}{t}\operatorname{cothc}_p(\sqrt[p]{C\mu}t))$ and $\beta(t) = \log(Ct^p \operatorname{sinhc}_p^p(\sqrt[p]{C\mu}t))$:

$$\begin{aligned} \dot{X} &= \frac{p}{t}\operatorname{cothc}_p(\sqrt[p]{C\mu}t)(Z - X) \\ \frac{d}{dt}\nabla h(Z) &= Cpt^{p-1}\operatorname{tanhc}_p^{p-1}(\sqrt[p]{C\mu}t) \\ &\quad \times \big(\mu\nabla h(X) - \mu\nabla h(Z) - \nabla f(X)\big), \end{aligned} \tag{12}$$

where C is a positive constant. We refer to this system of ODEs as the *unified ATM ODE*. Note that this system is equivalent to the unified AGM ODE (8) when $p = 2$, $C = 1/4$, and $h(x) = \frac{1}{2}\|x\|^2$. This system has a unique solution as shown in Appendix E.1. We address the convergence rate of this system in the following theorem.

Theorem 5.1. *The solution to the unified ATM ODE* (12) *with the initial conditions* $X(0) = Z(0) = x_0$ *satisfies*

$$\begin{aligned} f(X(t)) - f(x^*) &\le \frac{\operatorname{cschc}_p^p(\sqrt[p]{C\mu}t)}{Ct^p} D_h(x^*, x_0) \\ &= O\big(\min\{1/t^p, e^{-p\sqrt[p]{C\mu}t}\}\big). \end{aligned} \tag{13}$$

The proof of Theorem 5.1 can be found in Appendix E.2

5.2. Proposed Algorithm: Unified ATM

Recall that AGM is a combination of the gradient update step $x_{k+1} = y_k - s\nabla f(y_k)$ and the momentum steps. In order to extend AGM to the higher-order setting, we replace the gradient update step with its higher-order generalization, the *tensor update* step. For $p \ge 2$, $s > 0$, and $N > 0$, the *tensor update operator* $G_{p,s,N}$ is defined on $\mathbb{R}^n$ as

$$G_{p,s,N}(y) = \arg\min_x \left\{ f_{p-1}(x; y) + \frac{N}{ps}\|x - y\|^p \right\},$$

where $f_{p-1}(x; y) = \sum_{i=0}^{p-1} \frac{1}{i!}\nabla^i f(y)(x - y)^i$. As an easy consequence of (Wibisono et al., 2016, Lemma 2.2), we have the following lemma.

Lemma 5.2. *For any $p \ge 2$ and $N > 1$, there exists a positive constant M such that the inequality*

$$\langle \nabla f(x), y - x\rangle \ge Ms^{\frac{1}{p-1}} \|\nabla f(x)\|^{\frac{p}{p-1}} \tag{14}$$

holds for all $x, y \in \mathbb{R}^n$ with $x = G_{p,s,N}(y)$. In particular, for any $p \ge 2$ and $N = \sqrt{2}$, (14) holds with $M = 1/3$.

We now present the *unified ATM*, a rate-matching discretization of the unified ATM ODE (12), in Algorithm 2. The convergence rate of this algorithm is shown in the following theorem.

Algorithm 2 Unified ATM

Input: Initial point $x_0 \in \mathbb{R}^n$, stepsize s, positive constants N and M satisfying (14)
Initialize $z_0 = x_0$ and $A_0 = 0$; $C = \frac{1}{p}(\frac{M}{p-1})^{p-1}$
for $k = 0, 1, 2, \ldots$ **do**
  if $k = 0$ **then** $A_{k+1} = Cp^p s$
  else $A_{k+1} = A_k + p\sqrt[p]{CsA_k^{p-1}(1 + \mu A_k)}$
  $y_k = x_k + \frac{A_{k+1} - A_k}{A_{k+1}}(z_k - x_k)$
  $x_{k+1} = G_{p,s,N}(y_k)$
  $z_{k+1} = \arg\min_z \big\{ \frac{A_{k+1} - A_k}{1 + \mu A_k}\big(\langle \nabla f(x_{k+1}), z\rangle + \mu D_h(z, x_{k+1})\big) + D_h(z, z_k)\big\}$
end for

Theorem 5.3. *The iterates of the unified ATM (Algorithm 2) with $s \le (p-1)!/L$ satisfy*

$$\begin{aligned} f(x_k) - f(x^*) &\le \frac{1}{A_k} D_h(x^*, x_0) \\ &= O\big(\min\{1/k^p, (1 + p\sqrt[p]{C\mu s})^{-k}\}\big). \end{aligned} \tag{15}$$

The proof of Theorem 5.3 can be found in Appendix E.3. In particular, we can show that the unified ATM with $N = \sqrt{2}$, $M = 1/3$, and $s = \frac{(p-1)!}{L}$ has an $O(\sqrt[p]{L/\mu}\log(1/\epsilon))$ iteration complexity to find an ϵ-approximate solution,[9] which is equivalent to the iteration complexity of the restarted ATM-C (Wibisono et al., 2016, Appendix H), a known accelerated tensor method for uniformly convex objective functions.

[6] $h(y) \ge h(x) + \langle \nabla h(x), y - x\rangle + \frac{1}{p}\|y - x\|^p$ for all $x, y \in \mathbb{R}^n$.
[7] $\|\nabla^{p-1} f(y) - \nabla^{p-1} f(x)\| \le L\|y - x\|$ for all $x, y \in \mathbb{R}^n$.
[8] $\mu D_h(x, y) \le D_f(x, y)$ for all $x, y \in \mathbb{R}^n$.
[9] This follows from the fact that the inequality $(1 + p\sqrt[p]{C\mu s})^{-k} \le \exp(-\frac{1}{9}\sqrt[p]{\mu s}k)$ holds when $N = \sqrt{2}$ and $M = 1/3$ (see Remark E.1).

Algorithm 3 ATM-SC

Input: Initial point $x_0 \in \mathbb{R}^n$, stepsize s, positive constants N and M satisfying (14)
Initialize $z_0 = x_0$; $q = \mu s$, $C = \frac{1}{p}(\frac{M}{p-1})^{p-1}$
for $k = 0, 1, 2, \dots$ **do**
$y_k = x_k + \frac{p\sqrt[p]{C\mu s}}{1+p\sqrt[p]{C\mu s}}(z_k - x_k)$
$x_{k+1} = G_{p,s,N}(y_k)$
$z_{k+1} = \arg\min_z \Big\{ \frac{p\sqrt[p]{C\mu s}}{\mu}\big(\langle \nabla f(x_{k+1}), z\rangle + \mu D_h(z, x_{k+1})\big) + D_h(z, z_k)\Big\}$
end for

Unified ATM converges to the unified ATM ODE. In Appendix E.4, we show that the iterates of the unified ATM converge to the solution to the unified ATM ODE under the identifications $t \leftrightarrow \mathbf{t}_k$, $X(\mathbf{t}_k) \leftrightarrow x_k$, and $Z(\mathbf{t}_k) \leftrightarrow z_k$, where $\mathbf{t}_k := \sqrt[p]{A_k/C}$ if $\mu = 0$ and $\mathbf{t}_k := \sinh_p^{-1}(\sqrt[p]{\mu A_k})/\sqrt[p]{C\mu}$ if $\mu > 0$. Because the convergence rates (13) and (15) are equivalent under these identifications, the unified ATM is a rate-matching discretization of the unified ATM ODE.

5.3. Limit Cases of Unified ATM ODE and Unified ATM

We now examine two limit cases of the proposed dynamics and algorithm. First, when $\mu = 0$, the unified ATM ODE and the (modified) unified ATM reduce to ATM-C ODE and ATM-C in (Wibisono et al., 2016), respectively (see Ramerk E.2). Second, when $\mu > 0$, by taking the limits $t_0 \to \infty$ and $k_0 \to \infty$, we obtain novel dynamics and algorithm for minimizing uniformly convex functions. In Appendix E.5, we show that the asymptotic limit of the unified ATM ODE is given by

$$\dot{X} = p\sqrt[p]{C\mu}(Z - X)$$
$$\frac{d}{dt}\nabla h(Z) = p\sqrt[p]{C\mu}\left(\nabla h(X) - \nabla h(Z) - \frac{1}{\mu}\nabla f(X)\right)$$

and that the solution to this system achieves an $O(e^{-p\sqrt[p]{C\mu}t})$ convergence rate. In Appendix E.6, we derive the asymptotic limit of the unified ATM, resulting in *ATM-SC* (Algorithm 3), a time invariant method achieving an $O((1 + p\sqrt[p]{C\mu s})^{-k})$ convergence rate. As expected, these dynamics and algorithm are only applicable to the uniformly convex case ($\mu > 0$).

6. ODE Model for Minimizing the Gradient Norm of Strongly Convex Functions

Until now, we have focused on the dynamics that minimize the objective function value $f(X(t))$. In certain cases, the squared gradient norm $\|\nabla f(X(t))\|^2$ is also a reasonable performance measure for both theoretical and practical purposes (see Nesterov, 2012; Diakonikolas & Wang, 2022). In this section, we propose a novel ODE model which reduces the squared gradient norm of strongly convex functions with an $O\big(\min\{1/T^2, e^{-\sqrt{\mu}T}\}\big)$ convergence rate.

6.1. Motivation: Anti-Transpose Relationship Between OGM ODE and OGM-G ODE

To guide the design of our novel ODE model, we first investigate a symmetric relationship between

$$\ddot{X} + \frac{3}{t}\dot{X} + 2\nabla f(X) = 0, \qquad \text{(OGM ODE)}$$

an ODE that reduces the objective function value accuracy of convex functions, and

$$\ddot{X} + \frac{3}{T-t}\dot{X} + 2\nabla f(X) = 0, \qquad \text{(OGM-G ODE)}$$

an ODE that reduces the squared gradient norm of convex functions (see Appendix F.1 for details about these dynamics).

Based on the observation of a symmetric relationship between the coefficients of $\dot{X}$ in the two ODEs, one might guess that "*OGM-G ODE is the time-reversed dynamics of OGM ODE*." However, this interpretation is misleading as the solution to the two ODEs do not have a time-reversed relationship. Instead, using the differential kernel (see Section 2.2), we reveal a conceivably more accurate symmetric relationship between these ODEs.

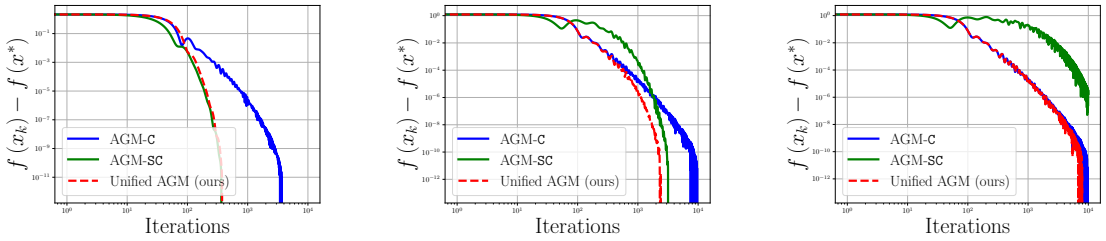
The differential kernels $H^{\mathrm{F}}(t,\tau)$ of OGM ODE and $H^{\mathrm{G}}(t,\tau)$ of OGM-G ODE can be computed as follows: $H^{\mathrm{F}}(t,\tau) = 2\tau^3/t^3$, and $H^{\mathrm{G}}(t,\tau) = 2(T-t)^3/(T-\tau)^3$, respectively. Here, we can observe the following *anti-transpose* relationship between the two differential kernels:[10]

$$H^{\mathrm{F}}(t,\tau) = H^{\mathrm{G}}(T-\tau, T-t). \qquad (16)$$

6.2. Proposed Dynamics: Unified AGM-G ODE

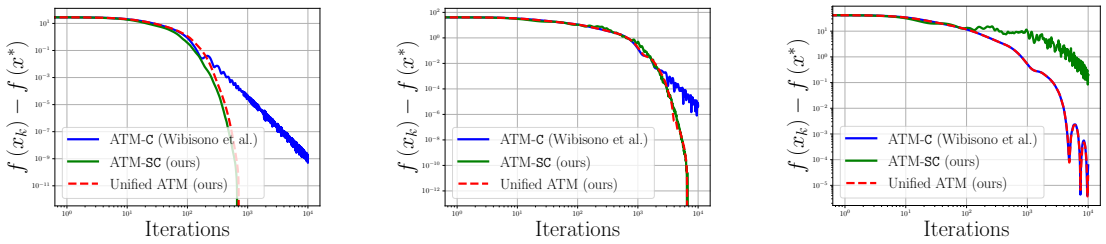
Suh et al. (2022) showed that OGM-G ODE reduces the squared gradient norm $\|\nabla f(X(T))\|^2$ at the terminal time T, with an $O(1/T^2)$ convergence rate. However, as OGM-G ODE exploits only the non-strong convexity of f, it cannot attain an exponential convergence rate. To overcome this limitation, we propose a novel ODE model that fully exploits the strong convexity of f. Inspired by the anti-transpose relationship between OGM ODE and OGM-G ODE, we

[10] In Appendix F.2, we show that this relationship can also be derived from the *anti-transpose* relationship (111) between the discrete-time algorithms OGM and OGM-G.



(a) $\lambda = 5$ ($\mu = 10^{-1}$) (b) $\lambda = 5 \cdot 10^{-2}$ ($\mu = 10^{-3}$) (c) $\lambda = 5 \cdot 10^{-4}$ ($\mu = 10^{-5}$)

Figure 2. Results for the ℓ_2-regularized logistic regression problem.



(a) $\lambda = 1$ ($\mu = 3 \cdot 10^{-3}$) (b) $\lambda = 10^{-3}$ ($\mu = 3 \cdot 10^{-6}$) (c) $\lambda = 10^{-6}$ ($\mu = 3 \cdot 10^{-9}$)

Figure 3. Results for the cubic-regularized linear regression problem.

propose the *unified AGM-G ODE* as

$$\ddot{X} + \left(\frac{\sqrt{\mu}}{2} \tanh\left(\frac{\sqrt{\mu}}{2}(T-t) \right) + \frac{3}{T-t} \times \operatorname{cothc}\left(\frac{\sqrt{\mu}}{2}(T-t) \right) \right) \dot{X} + \nabla f(X) = 0, \quad (17)$$

which is the *anti-transposed dynamics* of the unified AGM ODE (see Appendix F.3). The following theorem shows that this ODE model reduces the gradient norm of strongly convex functions at both polynomial and exponential rates.

Theorem 6.1. *The solution to the unified AGM-G ODE* (17) *with the initial conditions* $X(0) = x_0$ *and* $\dot{X}(0) = 0$ *satisfies*[11]

$$\|\nabla f(X(T))\|^2 \le \frac{8}{T^2} \operatorname{cschc}^2\left(\frac{\sqrt{\mu}}{2} T \right) \times \sup_x \left\{ f(x_0) - f(x) + \frac{\mu}{2}\|x_0 - x\|^2 \right\} = O\big(\min\{1/T^2, e^{-\sqrt{\mu}T}\} \big).$$

The proof of Theorem 6.1 can be found in Appendix F.4. In Remark F.1, we discuss that our unified framework is crucial for designing ODE models that reduces the gradient norm of strongly convex functions, as the proof of Theorem 6.1 relies on the property $\dot{X}(T) = 0$, which does not hold for the *anti-transposed dynamics* of AGM-SC ODE.

7. Numerical Experiments

In this section, we empirically test the performances of our unified algorithms (the unified AGM and unified ATM) against the specialized algorithms for non-strongly convex objective functions (AGM-C and ATM-C) and the specialized algorithms for strongly convex objective functions (AGM-SC and ATM-SC).

ℓ_2-regularized logistic regression. Consider the problem (1) with the convex functions $f_i(x) = -y_i a_i^T x + \log(1 + e^{a_i^T x})$ and the ℓ_2-regularization term $R(x) = \|x\|^2$, that is,

$$\min_{x \in \mathbb{R}^n} f(x) = \frac{1}{m}\left(\sum_{i=1}^m (-y_i a_i^T x + \log(1 + e^{a_i^T x})) + \lambda \|x\|^2 \right),$$

where $a_i \in \mathbb{R}^n$ and $y_i \in \{0, 1\}$. Then, the function f is $\frac{2\lambda}{m}$-strongly convex. We set the parameters as $s = 10^{-2}$ (step-size), $m = 100$, and $n = 20$. The vectors a_i and y_i were synthetically generated.[12] As shown in Figure 2, AGM-SC outperforms AGM-C when μ is large, but underperforms

[11] Here, we assume that $\sup_x \{f(x_0) - f(x) + \frac{\mu}{2}\|x_0 - x\|^2\}$ is finite. This assumption is quite mild because the function $x \mapsto f(x_0) - f(x) + \frac{\mu}{2}\|x_0 - x\|^2$ is concave when f is μ-strongly convex.

[12] The entries of a_i were sampled from $\mathcal{N}(0, 1)$, the labels $y_i \in \{0, 1\}$ were generated using the logistic model $P(y_i = 1) = 1/(1 + e^{-a_i^T x^0})$, and the entries of $x^0 \in \mathbb{R}^n$ were sampled from $\mathcal{N}(0, 1/100)$.

AGM-C when μ is small. In both cases, the performance of the unified AGM (Algorithm 1) is comparable to the better method among AGM-C and AGM-SC.

Cubic-regularized linear regression. To validate the performance of higher-order methods, we consider the problem

$$\min_{x\in\mathbb{R}^n} f(x) = \frac{1}{n}\left(\frac{1}{2}\|Ax-b\|^2 + \lambda\|x\|^3\right),$$

where $A \in \mathbb{R}^{n\times n}$ and $b \in \mathbb{R}^n$. Then, because $x \mapsto \|x\|^3$ is 6-smooth of order 2 and $\frac{3}{2}$-uniformly convex of order 3 (see Nesterov, 2008), the function f is $\frac{6\lambda}{n}$-smooth of order 2 and $\frac{3\lambda}{2n}$-strongly convex with respect to h, where $h(x) = \frac{2}{3}\|x\|^3$, which is an 1-uniformly convex function of order 3. We set the parameters as $s = 10^{-4}$ (stepsize), $p = 2$ (order), $N = \sqrt{2}$, $M = 1/3$ (input constants), and $n = 50$. The matrix A and the vector b were synthetically generated.[13] Figure 3 shows that the performance of the unified ATM (Algorithm 2) is comparable to the better method among ATM-C (Wibisono et al., 2016) and ATM-SC (Algorithm 3).

8. Concluding Remarks

We have developed a unified framework for designing accelerated continuous-time dynamics and discrete-time algorithms that handle convex and strongly convex functions simultaneously. Our unified framework has strong potential for future research since it resolves the inconsistencies that are commonly observed in the literature. On a different note, the newly proposed differential kernel may be improved in the future; for instance, it could be adapted to the mirror descent setup. Moreover, a rate-matching discretization of the unified AGM-G ODE could be further investigated.

Acknowledgements

This work was supported in part by Samsung Electronics, the National Research Foundation of Korea funded by MSIT(2020R1C1C1009766), and the Information and Communications Technology Planning and Evaluation (IITP) grant funded by MSIT(2022-0-00124, 2022-0-00480).

References

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, 2018.

Baes, M. Estimate sequence methods: extensions and approximations. *Institute for Operations Research, ETH, Zürich, Switzerland*, pp. 2, 2009.

[13] We set $A = B + B^{\mathrm{T}}$, where the entries of B were sampled from $\mathcal{N}(0,1)$. The entries of b were sampled from $\mathcal{N}(0,100)$.

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

Bubeck, S. et al. Convex optimization: Algorithms and complexity. *Foundations and Trends® in Machine Learning*, 8(3-4):231–357, 2015.

d'Aspremont, A., Scieur, D., and Taylor, A. Acceleration methods. *arXiv preprint arXiv:2101.09545*, 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.

Diakonikolas, J. and Wang, P. Potential function-based framework for minimizing gradients in convex and min-max optimization. *SIAM Journal on Optimization*, 32(3): 1668–1697, 2022.

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

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, 2021.

Gasnikov, A., Dvurechensky, P., Gorbunov, E., Vorontsova, E., Selikhanovych, D., and Uribe, C. A. Optimal tensor methods in smooth convex and uniformly convex-optimization. In *Conference on Learning Theory*, pp. 1374–1391. PMLR, 2019.

Ito, M. and Fukuda, M. Nearly optimal first-order methods for convex optimization under gradient norm measure: An adaptive regularization approach. *Journal of Optimization Theory and Applications*, 188:770–804, 2021.

Kim, D. and Fessler, J. A. Optimized first-order methods for smooth convex minimization. *Mathematical Programming*, 159(1):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. In *Advances in Neural Information Processing Systems*, volume 28, 2015.

Lee, J., Park, C., and Ryu, E. A geometric structure of acceleration and its role in making gradients small fast. *Advances in Neural Information Processing Systems*, 34: 11999–12012, 2021.

Luo, H. and Chen, L. From differential equation solvers to accelerated first-order methods for convex optimization. *Mathematical Programming*, pp. 1–47, 2021.

Lyapunov, A. M. The general problem of the stability of motion. *International Journal of Control*, 55(3):531–534, 1992.

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.

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

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

Nesterov, Y. *Lectures on Convex Optimization*, volume 137. Springer, 2018.

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

Rockafellar, R. T. and Wets, R. J.-B. *Variational Analysis*, volume 317. Springer Science & Business Media, 2009.

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

Shi, B., Du, S. S., Su, W., and Jordan, M. I. Acceleration via symplectic discretization of high-resolution differential equations. In *Advances in Neural Information Processing Systems*, volume 32, 2019.

Shi, B., Du, S. S., Jordan, M. I., and Su, W. J. Understanding the acceleration phenomenon via high-resolution differential equations. *Mathematical Programming*, pp. 1–70, 2021.

Su, W., Boyd, S., and Candes, E. A differential equation for modeling nesterov's accelerated gradient method: Theory and insights. In *Advances in Neural Information Processing Systems*, pp. 2510–2518, 2014.

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

Suh, J. J., Roh, G., and Ryu, E. K. Continuous-time analysis of accelerated gradient methods via conservation laws in dilated coordinate systems. In *International Conference on Machine Learning*, pp. 20640–20667. PMLR, 2022.

Taylor, A. and Drori, Y. An optimal gradient method for smooth strongly convex minimization. *Mathematical Programming*, pp. 1–38, 2022.

ten Thije Boonkkamp, J., van Dijk, J., Liu, L., and Peerenboom, K. S. Extension of the complete flux scheme to systems of conservation laws. *Journal of Scientific Computing*, 53(3):552–568, 2012.

Teschl, G. *Ordinary Differential Equations and Dynamical Systems*, volume 140. American Mathematical Soc., 2012.

Tseng, P. On accelerated proximal gradient methods for convex-concave optimization. *submitted to SIAM Journal on Optimization*, 2(3), 2008.

Wibisono, A., Wilson, A. C., and Jordan, M. I. A variational perspective on accelerated methods in optimization. *arXiv preprint arXiv:1603.04245*, 2016.

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, P., Orvieto, A., Daneshmand, H., Hofmann, T., and Smith, R. S. Revisiting the role of euler numerical integration on acceleration and stability in convex optimization. In *International Conference on Artificial Intelligence and Statistics*, pp. 3979–3987. PMLR, 2021.

A. Related Work

A.1. Inconsistencies Between ODE Models and Lagrangian Formulations

In this section, we discuss the inconsistencies inherent in ODE models and Lagrangian formulations for the two algorithms AGM-C and AGM-SC.

A.1.1. Inconsistency Between ODE Models

Limiting ODE of AGM-C. Recall that AGM-C is the three-sequence scheme (41) with $\tau_k = \frac{2}{k+1}$ and $\delta_k = \frac{s(k+1)}{2}$. With the sequence $\mathbf{t}_k = k\sqrt{s}$, we have

$$\tau(t) = \lim_{s\to 0} \frac{\tau_{\mathbf{k}(t)}}{\sqrt{s}} = \lim_{s\to 0} \frac{2}{\sqrt{s}\,(t/\sqrt{s}+1)} = \frac{2}{t}$$
$$\delta(t) = \lim_{s\to 0} \frac{\delta_{\mathbf{k}(t)}}{\sqrt{s}} = \lim_{s\to 0} \frac{\sqrt{s}\,(t/\sqrt{s}+1)}{2} = \frac{t}{2}.$$

Thus, as $s \to 0$, AGM-C converges to the following ODE system, which we call AGM-C system:

$$\begin{aligned} \dot{X} &= \frac{2}{t}(Z - X) \\ \dot{Z} &= -\frac{t}{2}\nabla f(X) \end{aligned} \tag{18}$$

with $X(0) = Z(0) = x_0$. This system can be written in the following second-order ODE, which we call AGM-C ODE:

$$\ddot{X} + \frac{3}{t}\dot{X} + \nabla f(X) = 0 \tag{19}$$

with $X(0) = x_0$ and $\dot{X}(0) = 0$. Su et al. (2014) first derived this ODE and showed that the solution to AGM-C ODE satisfies an $O(\|x_0 - x^*\|^2/t^2)$ convergence rate.

Limiting ODE of AGM-SC. Recall that AGM-SC is the three-sequence scheme (41) with $\tau_k = \frac{\sqrt{\mu s}}{1+\sqrt{\mu s}}$ and $\delta_k = \sqrt{\frac{s}{\mu}}$. With the sequence $\mathbf{t}_k = -k\frac{\log(1-\sqrt{\mu s})}{\sqrt{\mu}}$,[14] we have

$$\tau(t) = \lim_{s\to 0} \frac{\tau_{\mathbf{k}(t)}}{\sqrt{s}} = \lim_{s\to 0} \frac{\sqrt{\mu}}{1+\sqrt{\mu s}} = \sqrt{\mu}$$
$$\delta(t) = \lim_{s\to 0} \frac{\delta_{\mathbf{k}(t)}}{\sqrt{s}} = \lim_{s\to 0} \frac{1}{\sqrt{\mu}} = \frac{1}{\sqrt{\mu}}.$$

Thus, as $s \to 0$, AGM-SC converges to the following ODE system, which we call AGM-SC system:

$$\begin{aligned} \dot{X} &= \sqrt{\mu}(Z - X) \\ \dot{Z} &= \frac{1}{\sqrt{\mu}}\left(\mu X - \mu Z - \nabla f(X)\right) \end{aligned} \tag{20}$$

with $X(0) = Z(0) = x_0$, or equivalently, the following AGM-SC ODE:

$$\ddot{X} + 2\sqrt{\mu}\dot{X} + \nabla f(X) = 0 \tag{21}$$

with $X(0) = x_0$ and $\dot{X}(0) = 0$. Wilson et al. (2021) showed that the solution to this ODE satisfies an $O(e^{-\sqrt{\mu}t}(f(x_0) - f(x^*) + \frac{\mu}{2}\|x_0 - x^*\|^2))$ convergence rate. Just like in the discrete-time case, AGM-C ODE and AGM-SC ODE should be handled as separate cases because AGM-SC ODE does not recover AGM-C ODE as $\mu \to 0$.

AGM-SC ODE does not recover AGM-C ODE as $\mu \to 0$.

[14] Although the sequence $\mathbf{t}_k = k\sqrt{s}$ leads to the same limiting dynamics, this particular sequence makes a clear connection between the convergence rates of AGM-SC and its limiting ODE, as both rates are equivalent if we identify $t \leftrightarrow \mathbf{t}_k$, $X(\mathbf{t}_k) \leftrightarrow x_k$, and $Z(\mathbf{t}_k) \leftrightarrow z_k$.

A.1.2. Inconsistency Between Lagrangian Formulations

To systematically study the acceleration phenomenon of momentum methods, Wibisono et al. (2016) introduced the following *first* Bregman Lagrangian:

$$\mathcal{L}_{\text{1st}}\left(X, \dot{X}, t\right) = e^{\alpha+\gamma}\left(D_h\left(X + e^{-\alpha}\dot{X}, X\right) - e^{\beta} f(X)\right), \tag{22}$$

where $\alpha, \beta, \gamma : [0, \infty) \to \mathbb{R}$ are continuously differentiable functions, h is a continuously differentiable strictly convex function, and D_h is the Bregman divergence. In order to obtain accelerated convergence rates, the following *ideal scaling conditions* are introduced:

$$\dot{\gamma} = e^{\alpha} \tag{23a}$$

$$\dot{\beta} \leq e^{\alpha}. \tag{23b}$$

Under the ideal scaling condition (23a), the Euler–Lagrange equation

$$\frac{d}{dt}\left\{\frac{\partial \mathcal{L}}{\partial \dot{X}}\left(X, \dot{X}, t\right)\right\} = \frac{\partial \mathcal{L}}{\partial X}\left(X, \dot{X}, t\right) \tag{24}$$

for the first Bregman Lagrangian (22) reduces to the following system of first-order equations:

$$\dot{X} = e^{\alpha}(Z - X) \tag{25a}$$

$$\frac{d}{dt}\nabla h(Z) = -e^{\alpha+\beta}\nabla f(X). \tag{25b}$$

When f is convex, any solution to the system of ODEs (25) reduces the objective function value accuracy at an $O(e^{-\beta(t)})$ convergence rate. In particular, setting $\alpha(t) = \log \frac{2}{t}$ and $\beta(t) = \log \frac{t^2}{4}$, we recover AGM-C system (18) and its convergence rate.

Although the first Bregman Lagrangian (22) generates a large family of momentum dynamics, it does not include AGM-SC system (20). To handle strongly convex cases, Wilson et al. (2021) introduced the *second* Bregman Lagrangian, defined as

$$\mathcal{L}_{\text{2nd}}\left(X, \dot{X}, t\right) = e^{\alpha+\beta+\gamma}\left(\mu D_h\left(X + e^{-\alpha}\dot{X}, X\right) - f(X)\right). \tag{26}$$

Under the ideal scaling condition (23a), the Euler–Lagrange equation (24) for the second Bregman Lagrangian (26) reduces to the following system of first-order equations:

$$\dot{X} = e^{\alpha}(Z - X) \tag{27a}$$

$$\frac{d}{dt}\nabla h(Z) = \dot{\beta}\left(\nabla h(X) - \nabla h(Z)\right) - \frac{e^{\alpha}}{\mu}\nabla f(X). \tag{27b}$$

When f is μ-uniformly convex with respect to h, any solution to the system of ODEs (27) satisfies an $O(e^{-\beta(t)})$ convergence rate. In particular, letting $\alpha(t) = \log \sqrt{\mu}$ and $\beta(t) = \sqrt{\mu}t$, we recover AGM-SC system (20) and its convergence rate. Here, we observe an inconsistency between the two Bregman Lagrangians.

The second Bregman Lagrangian does not recover the first Bregman Lagrangian as $\mu \to 0$.

A.2. Lyapunov Arguments for Convergence Analyses

A popular method for proving the convergence rates of momentum dynamics and algorithms is constructing an energy function non-increasing over time, called the Lyapunov function (Lyapunov, 1992). The particular analyses presented in this section handle discrete-time algorithms and the corresponding continuous-time dynamics using a single Lyapunov function, as in (Krichene et al., 2015). To prove the convergence rates of the given algorithm and associated dynamics, we take the following steps:

1. Define a time-dependent Lyapunov function $V : \mathbb{R}^n \times \mathbb{R}^n \times [0, \infty) \to [0, \infty)$.

2. Show that the continuous-time energy functional $\mathcal{E}(t) = V(X(t), Z(t), t)$ is monotonically non-increasing along the solution trajectory $(X, Z) : [0, \infty) \to \mathbb{R}^n \times \mathbb{R}^n$ of the ODE system.
3. Show that the discrete-time energy functional $\mathcal{E}_k = V(x_k, z_k, \mathbf{t}_k)$ is monotonically non-increasing along the iterates $(x_k, z_k) : \{0, 1, 2, \ldots\} \to \mathbb{R}^n \times \mathbb{R}^n$ of the algorithm.

The remainder of this subsection shows how we can apply this strategy to known algorithms.

AGM-C and AGM-C ODE. We define a time-dependent Lyapunov function as

$$V(X, Z, t) := \frac{1}{2} \|Z - x^*\|^2 + \frac{t^2}{4} \left(f(X) - f\left(x^*\right)\right). \tag{28}$$

Then, the continuous-time energy functional

$$\mathcal{E}(t) = V(X(t), Z(t), t) = \frac{1}{2} \|Z(t) - x^*\|^2 + \frac{t^2}{4} \left(f(X(t)) - f\left(x^*\right)\right)$$

is monotonically non-increasing along the solution trajectory of AGM-C ODE (18) (see Su et al., 2016). Writing $\mathcal{E}(t) \leq \mathcal{E}(0)$ explicitly, we obtain an $O(1/t^2)$ convergence rate as

$$f(X(t)) - f(x^*) \leq \frac{4}{t^2} \mathcal{E}(t) \leq \frac{4}{t^2} \mathcal{E}(0) = \frac{2}{t^2} \|x_0 - x^*\|^2.$$

For the iterates of AGM-C, the discrete-time energy function

$$\mathcal{E}_k = V(x_k, z_k, \mathbf{t}_k) = \frac{1}{2} \|z_k - x^*\|^2 + \frac{sk^2}{4} \left(f(x_k) - f\left(x^*\right)\right), \tag{29}$$

where $\mathbf{t}_k = k\sqrt{s}$, is monotonically non-increasing (see Ryu & Yin, 2022, Chapter 12). Hence, we obtain an $O(1/k^2)$ convergence rate.

AGM-SC and AGM-SC ODE. We define a time-dependent Lyapunov function as

$$V(X, Z, t) := e^{\sqrt{\mu} t} \left(\frac{\mu}{2} \|Z - x^*\|^2 + f(X) - f\left(x^*\right)\right). \tag{30}$$

Then we can show that AGM-SC ODE (20) achieves an $O(e^{-\sqrt{\mu} t})$ convergence rate by showing that the energy functional

$$\mathcal{E}(t) = V(X(t), Z(t), t) = e^{\sqrt{\mu} t} \left(\frac{\mu}{2} \|Z(t) - x^*\|^2 + f(X(t)) - f\left(x^*\right)\right)$$

is monotonically non-increasing along the solution trajectory of AGM-SC ODE (see Wilson et al., 2021). Similarly, we can show that AGM-SC achieves an $O((1 - \sqrt{\mu s})^k)$ convergence rate by showing that the energy functional

$$\mathcal{E}_k = V(x_k, z_k, \mathbf{t}_k) = (1 - \sqrt{\mu s})^{-k} \left(\frac{\mu}{2} \|z_k - x^*\|^2 + f(x_k) - f\left(x^*\right)\right),$$

where $\mathbf{t}_k = -k \frac{\log(1 - \sqrt{\mu s})}{\sqrt{\mu}}$, is non-increasing along the iterates of AGM-SC (see d'Aspremont et al., 2021, Section 4.5).

Bregman Lagrangians. We can show that the first Bregman Lagrangian flow (25) and the second Bregman Lagrangian flow (27) achieve an $O(e^{-\beta(t)})$ convergence rate by showing that the energy functional $\mathcal{E}(t) = V(X(t), Z(t), t)$ is monotonically non-increasing. For the first Bregman Lagrangian flow, the Lyapunov function V is defined as

$$V_{\mathrm{1st}}(X, Z, t) := D_h\left(x^*, Z\right) + e^{\beta(t)} \left(f(X) - f\left(x^*\right)\right). \tag{31}$$

Thus, we have the energy function

$$\mathcal{E}_{\mathrm{1st}}(t) := D_h\left(x^*, Z(t)\right) + e^{\beta(t)} \left(f(X(t)) - f\left(x^*\right)\right). \tag{32}$$

For the second Bregman Lagrangian flow, the Lyapunov function V is defined as

$$V_{\mathrm{2nd}}(X, Z, t) := e^{\beta(t)} \left(\mu D_h\left(x^*, Z\right) + f(X) - f\left(x^*\right)\right). \tag{33}$$

Thus, we have the energy function

$$\mathcal{E}_{\mathrm{2nd}}(t) := e^{\beta(t)} \left(\mu D_h\left(x^*, Z(t)\right) + f(X(t)) - f\left(x^*\right)\right). \tag{34}$$

See (Wibisono et al., 2016; Wilson et al., 2021) for the proofs.

A.3. Original NAG Flow

Luo & Chen (2021) designed the following ODE model for the *constant step scheme I* (Nesterov, 2018, Equation 2.2.19), which we call the *original NAG system*:

$$\begin{aligned} \dot{\gamma} &= \mu - \gamma \\ \dot{X} &= Z - X \\ \dot{Z} &= \frac{1}{\gamma}(\mu X - \mu Z - \nabla f(X)) \end{aligned} \tag{35}$$

with $X(0) = Z(0) = x_0$ and $\gamma(0) = \gamma_0 > 0$. Luo & Chen (2021, Section 6.2) showed that the *constant step scheme I* (Nesterov, 2018, Equation 2.2.19) can be viewed as a discretization of this ODE system with the timestep α_i, which is inductively defined in (94).

Using time rescaling technique, Luo & Chen (2021) also proposed the following system of ODEs (although most of their results directly deal with Equation (35)):

$$\begin{aligned} \dot{X}(t) &= a(t)(Z(t) - X(t)) \\ b(t)\dot{Z}(t) &= a(t)(\mu X(t) - \mu Z(t) - \nabla f(X(t))), \end{aligned} \tag{36}$$

where $a : [0, \infty) \to [0, \infty)$ is an arbitrary function and

$$b(t) = \gamma \left(\int_0^t a(s)\, ds \right).$$

In Appendix A.3.1, we show that the original NAG flow (35) is a special case of the unified Bregman Lagrnagian flow (5). In Appendix A.3.2, we show that the rescaled original NAG flow (36) can be expressed as the unified Bregman Lagrangian flow. Conversely, the unified Bregman Lagrangian flow can be expressed as the rescaled original NAG flow if the ideal scaling condition (23b) holds with equality and the distance-generating function h is Euclidean ($h(x) = \frac{1}{2}\|x\|^2$). Therefore, our unified Bregman Lagrangian generates a strictly larger family compared to (36). To emphasize, only our family can deal with the non-Euclidean setup (mirror descent setup). In addition, the derivation of our unified family (5) is more constructive because it comes from a Lagrangian formulation, whereas Luo & Chen (2021) designed the family (36) through a heuristic speculation. In Appendix A.3.3, we observe that the rescaled original NAG flow with specific parameters is closely related to the unified AGM ODE (8).

A.3.1. Original NAG Flow is a Special Case of Unified Bregman Lagrangian Flow

Solving $\dot{\gamma} = \mu - \gamma$ gives $\gamma(t) = \mu + (\gamma_0 - \mu)e^{-t}$. Thus, the original NAG system (35) can be written as

$$\begin{aligned} \dot{X} &= Z - X \\ \dot{Z} &= \frac{e^{t - \log(\gamma_0 - \mu)}}{1 + \mu e^{t - \log(\gamma_0 - \mu)}}(\mu X - \mu Z - \nabla f(X)). \end{aligned}$$

This ODE system is equivalent to the unified Bregman Lagrangian flow (5) with $\alpha(t) = 0$, $\beta(t) = t - \log(\gamma_0 - \mu)$, and $h(x) = \frac{1}{2}\|x\|^2$. Therefore, the original NAG flow (35) is a special case of the unified Bregman Lagrangian flow.

A.3.2. Time Rescaling Applied to Original NAG Flow Gives a Subfamily of Unified Bregman Lagrangian Flow

Because the uified Bregman Lagrangian flow is closed under time-dilation (see Appendix C.2), the rescaled original NAG flow (36) is a subfamily of the unified Bregman Lagrangian flow, by construction. In this subsubsection, we confirm this fact again. In addition, we show that the rescaled original NAG flow corresponds to the unified Bregman Lagrangian flow with the condition $\dot{\beta} = e^{\alpha}$ (ideal scaling condition (23b) with equality) and the Euclidean distance-genrerating function $h(x) = \frac{1}{2}\|x\|^2$.

First, we show that the rescaled original NAG flow (36) can be expressed as the unified Bregman Lagrangian flow (5). Given the parameter function $a(t)$ and the constant γ_0 of the rescaled original NAG flow, we can write the functions $\gamma(t)$ and $b(t)$ involved in (35) and (36) as

$$\gamma(t) = \mu + (\gamma_0 - \mu)\, e^{-t}$$

$$b(t) = \mu + (\gamma_0 - \mu)\, e^{-\int_0^t a(s)\, ds}.$$

We define the functions $\alpha(t)$ and $\beta(t)$ as

$$\begin{aligned} \alpha(t) &= \log a(t) \\ \beta(t) &= \log\left(\frac{1}{\gamma_0 - \mu}\right) + \int_0^t a(s)\, ds. \end{aligned} \tag{37}$$

Then, we have

$$\frac{\dot\beta e^\beta}{1 + \mu e^\beta} = \frac{e^{\alpha+\beta}}{1 + \mu e^\beta} = \frac{e^\alpha}{\mu + e^{-\beta}} = \frac{a(t)}{\mu + (\gamma_0 - \mu)\, e^{-\int_0^t a(s)\, ds}} = \frac{a(t)}{b(t)}.$$

Thus, the rescaled original NAG flow is equivalent to the unified Bregman Lagrangian flow with the parameter functions (37) and the Euclidean distance-generating function $h(x) = \frac{1}{2}\|x\|^2$.

Conversely, we show that if the ideal scaling conditon (23b) holds with equality and the distance-generating function h is Euclidean, then the unified Bregman Lagrangian flow can be written as the rescaled original NAG flow. Given the parameter functions $\alpha(t)$ and $\beta(t)$ of the unified Bregman Lagrangian flow, we define the function $a(t)$ and the constant γ_0 as

$$\begin{aligned} a(t) &= e^{\alpha(t)} \\ \gamma_0 &= \mu + e^{-\beta(0)}. \end{aligned}$$

Then, because

$$b(t) = \mu + (\gamma_0 - \mu)\, e^{-\int_0^t a(s)\, ds} = \mu + e^{-\beta(t)},$$

we can write the rescaled original NAG flow as

$$\begin{aligned} \dot X(t) &= e^{\alpha(t)}(Z(t) - X(t)) \\ \left(\mu + e^{-\beta(t)}\right)\dot Z(t) &= e^{\alpha(t)}(\mu X(t) - \mu Z(t) - \nabla f(X(t))), \end{aligned}$$

which is equivalent to the unified Bregman Lagrangian flow if the ideal scaling conditon (23b) holds with equality and $h(x) = \frac{1}{2}\|x\|^2$.

A.3.3. Relationship Between Original NAG Flow With Specific Parameters and Unified AGM ODE

As Luo & Chen (2021, Equation 70) mentioned, given $\gamma_0 > 0$, one can choose the function $a(t)$ in the rescaled original NAG flow as

$$a(t) = \begin{cases} \frac{2\sqrt{\gamma_0}}{\sqrt{\gamma_0}t + 2}, & \text{if } \mu = 0, \\ \sqrt{\mu} \cdot \frac{e^{\sqrt{\mu}t} - \frac{\sqrt{\mu} - \sqrt{\gamma_0}}{\sqrt{\mu} + \sqrt{\gamma_0}}}{e^{\sqrt{\mu}t} + \frac{\sqrt{\mu} - \sqrt{\gamma_0}}{\sqrt{\mu} + \sqrt{\gamma_0}}}, & \text{if } \mu > 0. \end{cases} \tag{38}$$

In this case, we have $b(t) = (a(t))^2$. Thus, the rescaled original flow with these functions can be written as

$$\begin{aligned} \dot X(t) &= \frac{2\sqrt{\gamma_0}}{\sqrt{\gamma_0}t + 2}(Z(t) - X(t)) \\ \dot Z(t) &= -\frac{\sqrt{\gamma_0}t + 2}{2\sqrt{\gamma_0}} - \nabla f(X(t)) \end{aligned}$$

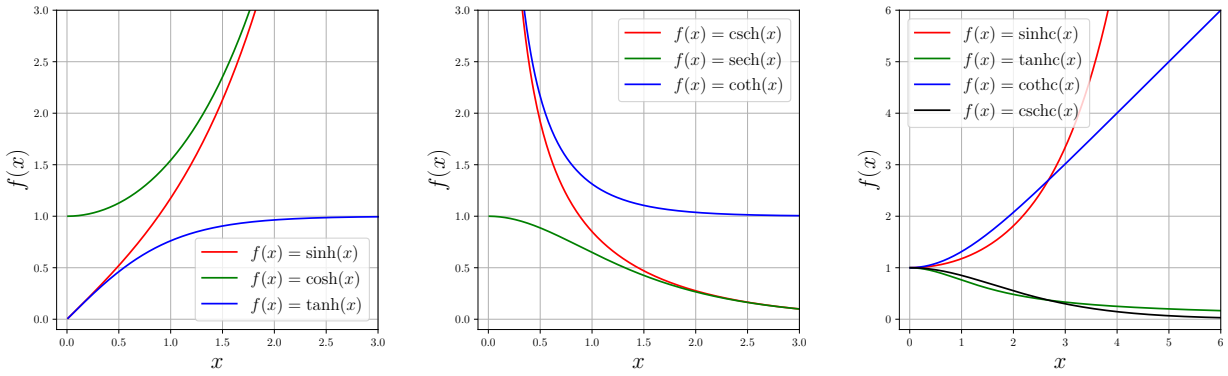
when $\mu = 0$, and

$$\begin{aligned} \dot X(t) &= \sqrt{\mu} \cdot \frac{e^{\sqrt{\mu}t} - \frac{\sqrt{\mu} - \sqrt{\gamma_0}}{\sqrt{\mu} + \sqrt{\gamma_0}}}{e^{\sqrt{\mu}t} + \frac{\sqrt{\mu} - \sqrt{\gamma_0}}{\sqrt{\mu} + \sqrt{\gamma_0}}}(Z(t) - X(t)) \\ \dot Z(t) &= \frac{1}{\sqrt{\mu}} \cdot \frac{e^{\sqrt{\mu}t} + \frac{\sqrt{\mu} - \sqrt{\gamma_0}}{\sqrt{\mu} + \sqrt{\gamma_0}}}{e^{\sqrt{\mu}t} - \frac{\sqrt{\mu} - \sqrt{\gamma_0}}{\sqrt{\mu} + \sqrt{\gamma_0}}}(\mu X(t) - \mu Z(t) - \nabla f(X(t))) \end{aligned}$$

when $\mu > 0$. Because $\frac{\sqrt{\mu} - \sqrt{\gamma_0}}{\sqrt{\mu} + \sqrt{\gamma_0}} \to -1$ as $\gamma_0 \to \infty$ and $\frac{e^{\sqrt{\mu}t} + 1}{e^{\sqrt{\mu}t} - 1} = \coth(\frac{\sqrt{\mu}}{2}t)$, the rescaled original NAG flow with (38) converges to the unified AGM system (69) as $\gamma_0 \to \infty$.

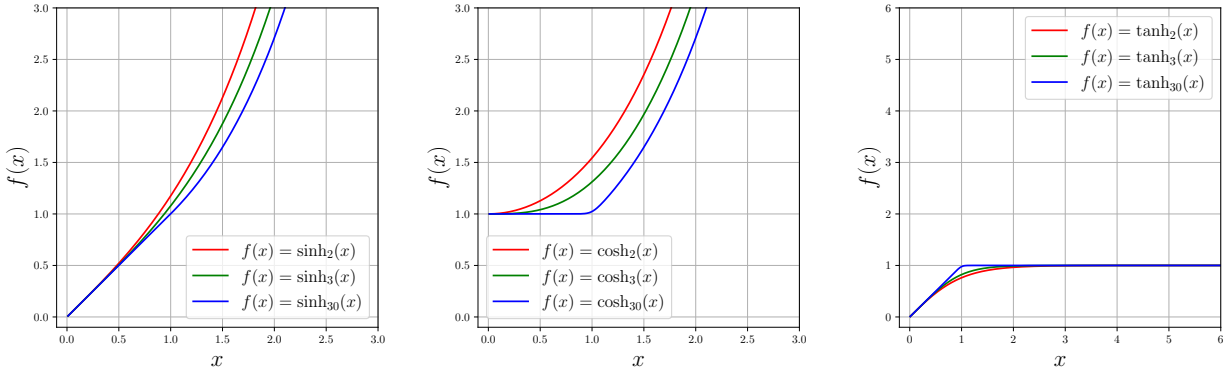
B. Preliminaries

B.1. Higher-Order Hyperbolic Functions



(a) sinh, cosh, tanh (b) coth, sech, csch (c) sinhc, tanhc, cothc, cschc

Figure 4. Hyperbolic functions and their variants.



(a) $\sinh_p$ (b) $\cosh_p$ (c) $\tanh_p$

Figure 5. Higher-order hyperbolic functions.

The following proposition indicates that the $\sinh_p$ function grows exponentially.

Proposition B.1. *There exists a constant $C_p > 0$ such that $\sinh_p(t) \sim C_p e^t$ as $t \to \infty$. In particular, we have $C_p = 1/2$ for $p = 2$.*

The proof of Proposition B.1 can be found in Appendix B.1.1. Using the definition of the $\sinh_p$ function and Proposition B.1, it is straightforward to check the following asymptotic properties:

$$\begin{aligned} \sinh_p x &\sim x \text{ as } x \to 0, & \sinh_p x &\sim C_p e^x \text{ as } x \to \infty \\ \cosh_p x &\sim 1 \text{ as } x \to 0, & \cosh_p x &\sim C_p e^x \text{ as } x \to \infty \\ \tanh_p x &\sim x \text{ as } x \to 0, & \tanh_p x &\sim 1 \text{ as } x \to \infty. \end{aligned}$$

B.1.1. Proof of Proposition B.1

Fix $T > 0$. We will show that

$$\log\left(\sinh_p(T+t)\right) - t \tag{39}$$

converges to some constant as $t \to \infty$. We can bound the derivative of (39) as

$$\begin{aligned}\frac{d}{dt}\left\{\log\left(\sinh_p(T+t)\right) - t\right\} &= \frac{\sinh_p'(T+t)}{\sinh_p(T+t)} - 1 \\ &= \frac{\cosh_p(T+t)}{\sinh_p(T+t)} - 1 \\ &= \left(1 + \frac{1}{\sinh_p^p(T+t)}\right)^{1/p} - 1 \\ &\in \left[0, \frac{1}{\sinh_p(T+t)}\right],\end{aligned}$$

where the last line follows from the fact that $1 \le (1+x)^{1/p} \le 1 + x^{1/p}$ holds for $x \ge 0$.[15] Thus, if the integral

$$\int_0^\infty \frac{1}{\sinh_p(T+t)}\, dt \tag{40}$$

is finite, then (39) converges to some constant because it is monotonically increasing and bounded above, and thus this completes the proof. To show that the integral (40) is finite, it is enough to show that the inequality

$$\sinh_p(T+t) \ge \sinh_p(T) e^t$$

holds for all $t \ge 0$. This can be shown by the following calculation:

$$\begin{aligned}\log\left(\sinh_p(T+t)\right) &= \log\left(\sinh_p(T)\right) + \int_0^t \frac{d}{ds}\left\{\log\left(\sinh_p(T+s)\right)\right\} ds \\ &= \log\left(\sinh_p(T)\right) + \int_0^t \frac{\sinh_p'(T+s)}{\sinh_p(T+s)}\, ds \\ &= \log\left(\sinh_p(T)\right) + \int_0^t \frac{\left(1 + \sinh_p^p(T+s)\right)^{1/p}}{\sinh_p(T+s)}\, ds \\ &\ge \log\left(\sinh_p(T)\right) + \int_0^t 1\, ds \\ &= \log\left(\sinh_p(T)\right) + t \\ &= \log\left(\sinh_p(T) e^t\right).\end{aligned}$$

B.1.2. The Function sinhc_p is Non-Decreasing

It is easy to see that $\sinh_p$ and $\cosh_p$ are increasing. Since

$$\begin{aligned}\tanh_p'(t) &= \frac{d}{dt}\left\{\frac{\sinh_p(t)}{\cosh_p(t)}\right\} \\ &= \frac{\sinh_p'(t)\cosh_p(t) - \cosh_p'(t)\sinh_p(t)}{\cosh_p^2(t)} \\ &\le \frac{\sinh_p'(t)\cosh_p(t)}{\cosh_p^2(t)} \\ &= 1,\end{aligned}$$

we have $\tanh_p(t) \le t$ for all $t \ge 0$. Now, we deduce that

$$\mathrm{sinhc}_p'(t) = \frac{d}{dt}\left\{\frac{\sinh_p(t)}{t}\right\}$$

[15] To check this basic inequality, one can consider the p-th power of each side.

$$
\begin{aligned}
&= \frac{t\sinh_p'(t) - \sinh_p(t)}{t^2} \\
&= \frac{t\cosh_p(t) - \sinh_p(t)}{t^2} \\
&= \frac{\cosh_p(t)}{t^2}\left(t - \tanh_p(t)\right) \\
&\geq 0,
\end{aligned}
$$

and thus sinhc is non-decreasing.

B.2. Limiting Arguments

In this subsection, we investigate the limiting ODE of *two-sequence scheme* and the limiting ODE of the fixed-step first-order scheme (2). The first approach is to write the algorithm as a two-sequence scheme and then derive the limiting ODE via the second-order Taylor series expansion. This argument frequently appears in the literature (see Su et al., 2016; Shi et al., 2021). The second approach, which is novel, is to express the algorithm using the *difference matrix* $\mathbf{H} = (h_{ij})$ and then derive the *differential kernel* $H(t,\tau)$ corresponding to the matrix (h_{ij}).

Furthermore, we show that the limiting ODE of two-sequence scheme recovers the limiting ODE of three-sequence scheme and that the limiting ODE of the fixed-step first-order scheme recovers the limiting ODE of the two-sequence scheme.

B.2.1. Limiting Argument for Three-Sequence Scheme

We informally derive the limiting ODE of the following three-sequence scheme:

$$
\begin{aligned}
y_k &= x_k + \tau_k\left(z_k - x_k\right) && \text{(41a)}\\
x_{k+1} &= y_k - s\nabla f\left(y_k\right) && \text{(41b)}\\
z_{k+1} &= z_k + \delta_k\left(\mu y_k - \mu z_k - \nabla f\left(y_k\right)\right). && \text{(41c)}
\end{aligned}
$$

To identify a discrete-time sequence $(x_k)_{k=0}^{\infty}$ with a continuous-time curve $X:[0,\infty)\to\mathbb{R}^n$, given the algorithmic stepsize s, we introduce a strictly increasing sequence $(\mathbf{t}_k)_{k=0}^{\infty}$ (depending on s) in $[0,\infty)$ and make the identification $X(\mathbf{t}_k) = x_k$. We denote the inverse of the sequence $\mathbf{t}:\{0,1,2,\ldots\}\to\mathbb{R}$ as $\mathbf{k}$, that is, $\mathbf{k}(\mathbf{t}_k) = k$ for all $k\geq 0$. For convenience, we extend the function $\mathbf{k}$ to a piecewise linear function defined on $[0,\infty)$.

We assume that

$$
\lim_{s\to 0}\mathbf{t}_0 = 0 \tag{42}
$$

and that the timesteps are asymptotically equivalent to $\sqrt{s}$ as $s\to 0$ in the sense that

$$
\lim_{s\to 0}\frac{\mathbf{t}_{\mathbf{k}(t)+1} - t}{\sqrt{s}} = 1 \quad \forall t\in(0,\infty). \tag{43}
$$

Note that the popular choice $\mathbf{t}_k = t_k := k\sqrt{s}$ (we will use the notation t_k for this specific sequence throughout the paper) used in (Su et al., 2014; Wibisono et al., 2016; Shi et al., 2021) satisfies these conditions.

For the iterates of three-sequence scheme (41), we have

$$
\begin{aligned}
\frac{x_{k+1} - x_k}{\sqrt{s}} &= \frac{\tau_k}{\sqrt{s}}\left(z_k - x_k\right) - \sqrt{s}\nabla f\left(y_k\right)\\
\frac{z_{k+1} - z_k}{\sqrt{s}} &= \frac{\delta_k}{\sqrt{s}}\left(\mu y_k - \mu z_k - \nabla f\left(y_k\right)\right).
\end{aligned}
$$

We introduce two sufficiently smooth curves $X, Z:[0,\infty)\to\mathbb{R}^n$ (possibly depending on s now) such that $X(t) = x_{\mathbf{k}(t)}$ and $Z(t) = z_{\mathbf{k}(t)}$. Since $\|x_{k+1} - y_k\| = o(\sqrt{s})$ and ∇f is Lipschitz continuous, we have

$$
\dot{X}(t) = \lim_{s\to 0}\frac{x_{\mathbf{k}(t)+1} - x_{\mathbf{k}(t)}}{t_{\mathbf{k}(t)+1} - t} = \lim_{s\to 0}\frac{x_{\mathbf{k}(t)+1} - x_{\mathbf{k}(t)}}{\sqrt{s}} = \lim_{s\to 0}\left\{\frac{\tau_{\mathbf{k}(t)}}{\sqrt{s}}\right\}\left(Z(t) - X(t)\right)
$$

$$\dot{Z}(t) = \lim_{s\to 0} \frac{z_{\mathbf{k}(t)+1} - z_{\mathbf{k}(t)}}{t_{\mathbf{k}(t)+1} - t} = \lim_{s\to 0} \frac{z_{\mathbf{k}(t)+1} - z_{\mathbf{k}(t)}}{\sqrt{s}} = \lim_{s\to 0} \left\{ \frac{\delta_{\mathbf{k}(t)}}{\sqrt{s}} \right\} (\mu X(t) - \mu Z(t) - \nabla f(X(t)))$$

for all $t > 0$. Thus, if the limits

$$\begin{aligned} \tau(t) &= \lim_{s\to 0} \frac{\tau_{\mathbf{k}(t)}}{\sqrt{s}} \\ \delta(t) &= \lim_{s\to 0} \frac{\delta_{\mathbf{k}(t)}}{\sqrt{s}} \end{aligned} \tag{44}$$

exist for all $t \in (0, \infty)$, then as $s \to 0$, the iterates generated by the three-sequence scheme (41) converge to a solution to the following system of ODEs:

$$\begin{aligned} \dot{X}(t) &= \tau(t)(Z(t) - X(t)) \\ \dot{Z}(t) &= \delta(t)(\mu X(t) - \mu Z(t) - \nabla f(X(t))) \end{aligned} \tag{45}$$

with the initial conditions $X(0) = Z(0) = x_0$. We can equivalently write this as the following second-order ODE:

$$\ddot{X} + \left(\tau(t) - \frac{\dot{\tau}(t)}{\tau(t)} + \mu\delta(t) \right) \dot{X} + \tau(t)\delta(t)\nabla f(X) = 0. \tag{46}$$

Furthermore, when the collinearity condition[16]

$$1 - \mu\delta_k - (1/s - \mu)\tau_k\delta_k = 0. \tag{47}$$

holds, we have

$$\delta(t) = \lim_{s\to 0} \frac{\delta_k}{\sqrt{s}} = \lim_{s\to 0} \frac{1}{\sqrt{s}\,(\mu + (1/s - \mu)\tau_k)} = \lim_{s\to 0} \frac{\sqrt{s}}{\mu s + (1 - \mu s)\tau_k} = \frac{1}{\tau(t)}. \tag{48}$$

B.2.2. Limiting Argument for Two-Sequence Scheme

We consider the following *two-sequence scheme*:

$$\begin{aligned} x_{k+1} &= y_k - s\nabla f\left(y_k\right) \\ y_{k+1} &= x_{k+1} + \beta_k\left(x_{k+1} - x_k\right) + \gamma_k\left(x_{k+1} - y_k\right). \end{aligned} \tag{49}$$

If we have

$$\lim_{s\to 0} \frac{1 - \beta_{t/\sqrt{s}}}{\sqrt{s}} = b(t) \text{ and } \lim_{s\to 0} \gamma_{t/\sqrt{s}} = c(t) \quad \forall t > 0 \tag{50}$$

for some smooth functions $b, c : (0, \infty) \to \mathbb{R}$, then we will see that under the identification $X(t_k) = x_k$ with $t_k = k\sqrt{s}$, the two-sequence scheme (49) converges to the ODE

$$\ddot{X}(t) + b(t)\dot{X}(t) + (1 + c(t))\nabla f(X(t)) = 0 \tag{51}$$

as $s \to 0$. For the iterates of the two-sequence scheme (49), we have

$$\begin{aligned} \frac{x_{k+1} - x_k}{\sqrt{s}} &= \frac{1}{\sqrt{s}}\left(y_k - s\nabla f\left(y_k\right) - x_k\right) \\ &= \frac{1}{\sqrt{s}}\left(\beta_{k-1}\left(x_k - x_{k-1}\right) + \gamma_k\left(x_k - y_{k-1}\right) - s\nabla f\left(y_k\right)\right) \\ &= \frac{1}{\sqrt{s}}\left(\beta_{k-1}\left(x_k - x_{k-1}\right) - s\gamma_k\nabla f\left(y_{k-1}\right) - s\nabla f\left(y_k\right)\right) \\ &= \beta_{k-1}\frac{x_k - x_{k-1}}{\sqrt{s}} - \sqrt{s}\gamma_k\nabla f\left(y_{k-1}\right) - \sqrt{s}\nabla f\left(y_k\right). \end{aligned}$$

[16] This condition ensures that the points x_k, x_{k+1}, z_{k+1} are collinear. Thus, one can write the updating rule for y_k as $y_{k+1} = x_{k+1} + \beta_k(x_{k+1} - x_k)$ for some $\beta_k \in \mathbb{R}$. This property provides a clear momentum effect: The point y_{k+1} is defined by adding a momentum term $\beta_k\left(x_{k+1} - x_k\right)$ to the previous point x_{k+1}. This property is useful when generalizing AGM methods to handle non-smooth terms (see d'Aspremont et al., 2021, Algorithm 20).

Using the Taylor expansions

$$\frac{x_{k+1}-x_k}{\sqrt{s}} = \dot{X}(t_k) + \frac{1}{2}\ddot{X}(t_k)\sqrt{s} + o\left(\sqrt{s}\right)$$
$$\frac{x_k - x_{k-1}}{\sqrt{s}} = \dot{X}(t_k) - \frac{1}{2}\ddot{X}(t_k)\sqrt{s} + o\left(\sqrt{s}\right),$$

we obtain

$$\dot{X}(t_k) + \frac{1}{2}\ddot{X}(t_k)\sqrt{s} + o\left(\sqrt{s}\right) = \beta_{k-1}\left(\dot{X}(t_k) - \frac{1}{2}\ddot{X}(t_k)\sqrt{s} + o\left(\sqrt{s}\right)\right) - \sqrt{s}\gamma_k\nabla f\left(y_{k-1}\right) - \sqrt{s}\nabla f\left(y_k\right).$$

It follows from $\|x_k - y_{k-1}\| = o(\sqrt{s})$ and the Lipschitz continuity of ∇f that

$$\sqrt{s}\nabla f\left(y_{k-1}\right) = \sqrt{s}\nabla f(X(t_k)) + o\left(\sqrt{s}\right)$$
$$\sqrt{s}\nabla f\left(y_k\right) = \sqrt{s}\nabla f\left(y_{k-1}\right) + o\left(\sqrt{s}\right) = \sqrt{s}\nabla f(X(t_k)) + o\left(\sqrt{s}\right).$$

Substituting these into the ODE yields

$$\frac{1+\beta_{k-1}}{2}\ddot{X}(t_k)\sqrt{s} + \left(1-\beta_{k-1}\right)\dot{X}(t_k) + \left(1+\gamma_k\right)\nabla f(X(t_k))\sqrt{s} + o\left(\sqrt{s}\right) = 0.$$

Dividing both sides by $\sqrt{s}$, substituting $k = t/\sqrt{s}$ and the limits (50), and then letting $s \to 0$, we obtain (note that $\beta_{t/\sqrt{s}-1} \to 1$ by Equation (50))

$$\ddot{X}(t) + b(t)\dot{X}(t) + (1+c(t))\nabla f(X(t)) = 0.$$

Recovering the limiting ODE of three-sequence scheme. We can write the three-sequence scheme (41) as the two-sequence scheme (49) with the following parameters (see Lee et al., 2021, Appendix B):

$$\begin{aligned}\beta_k &= \frac{\left(1-\tau_k\right)\tau_{k+1}\left(1-\mu\delta_k\right)}{\tau_k} \\ \gamma_k &= \frac{\tau_{k+1}\left((1/s-\mu)\delta_k\tau_k - 1 + \mu\delta_k\right)}{\tau_k}.\end{aligned} \tag{52}$$

Assume that the limits (44) with $\mathbf{t}_k = k\sqrt{s}$ exist. Then, it follows from the Taylor expansion that

$$\begin{aligned}\tau_k &= \tau\left(t_k\right)\sqrt{s} \\ \tau_{k+1} &= \tau\left(t_k\right)\sqrt{s} + \dot{\tau}\left(t_k\right)s + \sqrt{s}o\left(\sqrt{s}\right) \\ \delta_k &= \delta\left(t_k\right)\sqrt{s}.\end{aligned}$$

Thus, for the sequences (β_k) and (γ_k) in (52), we have

$$\begin{aligned}\frac{1-\beta_k}{\sqrt{s}} &= \frac{1}{\sqrt{s}}\left(1 - \left(1-\tau_k\right)\left(1-\mu\delta_k\right)\frac{\tau_{k+1}}{\tau_k}\right) \\ &= \frac{1}{\sqrt{s}}\left(1 - \left(1-\sqrt{s}\tau\left(t_k\right)\right)\left(1-\mu\sqrt{s}\delta\left(t_k\right)\right)\left(1 + \frac{\dot{\tau}\left(t_k\right)s + \sqrt{s}o\left(\sqrt{s}\right)}{\tau\left(t_k\right)\sqrt{s}}\right)\right) \\ &= \frac{1}{\sqrt{s}}\left(\sqrt{s}\tau\left(t_k\right) + \mu\sqrt{s}\delta\left(t_k\right) - \sqrt{s}\frac{\dot{\tau}\left(t_k\right)}{\tau\left(t_k\right)} + o\left(\sqrt{s}\right)\right) \\ &= \tau\left(t_k\right) + \mu\delta\left(t_k\right) - \frac{\dot{\tau}\left(t_k\right)}{\tau\left(t_k\right)} + \frac{o\left(\sqrt{s}\right)}{\sqrt{s}}\end{aligned}$$

and

$$\gamma_k = \frac{\tau_{k+1}}{\tau_k}\left((1/s-\mu)\delta_k\tau_k - 1 + \mu\delta_k\right)$$

$$= \left(1 + \frac{\dot{\tau}(t_k)\sqrt{s} + o(\sqrt{s})}{\tau(t_k)}\right)\left((1-\mu s)\delta(t_k)\tau(t_k) - 1 + \mu\sqrt{s}\delta(t_k)\right)$$
$$= \delta(t_k)\tau(t_k) - 1 + o(1).$$

Thus, we have

$$\lim_{s\to 0}\frac{1-\beta_{t/\sqrt{s}}}{\sqrt{s}} = \tau(t) + \mu\delta(t) - \frac{\dot{\tau}(t)}{\tau(t)}$$
$$\lim_{s\to 0}\gamma_{t/\sqrt{s}} = \tau(t)\delta(t) - 1.$$

Therefore, we recover the limiting ODE (46) of the three-sequence scheme. In particular, if the algorithmic parameters (τ_k) and (δ_k) satisfy the collinearity condition (47), then we have $\gamma_k = 0$ for all $k \geq 0$, and thus $c(t) = 0$.

Two-sequence form of AGM-C. Because AGM-C is the three-sequence scheme (41) with $\tau_k = \frac{2}{k+1}$, $\delta_k = \frac{s(k+1)}{2}$, and $\mu = 0$, we can rewrite it as the two-sequence scheme (49) with

$$\beta_k = \frac{\left(1 - \frac{2}{k+1}\right)\frac{2}{k+2}}{\frac{2}{k+1}} = \frac{k-1}{k+2}$$
$$\gamma_k = \frac{\frac{2}{k+2}\cdot\frac{s(k+1)}{2}}{s} - \frac{\frac{2}{k+2}}{\frac{2}{k+1}} = 0.$$

Thus, AGM-C converges to the ODE (51) with

$$b(t) = \lim_{s\to 0}\frac{1 - \frac{t/\sqrt{s}-1}{t/\sqrt{s}+2}}{\sqrt{s}} = \frac{3}{t}$$
$$c(t) = 0,$$

which recovers AGM-C ODE.

Two-sequence form of AGM-SC. Because AGM-SC is the three-sequence scheme (41) with $\tau_k = \frac{\sqrt{\mu s}}{1+\sqrt{\mu s}}$ and $\delta_k = \sqrt{\frac{s}{\mu}}$, it can be written as the two-sequence scheme (49) with

$$\beta_k = \frac{\left(1 - \frac{\sqrt{\mu s}}{1+\sqrt{\mu s}}\right)\frac{\sqrt{\mu s}}{1+\sqrt{\mu s}}\left(1 - \mu\sqrt{\frac{s}{\mu}}\right)}{\frac{\sqrt{\mu s}}{1+\sqrt{\mu s}}} = \frac{1-\sqrt{\mu s}}{1+\sqrt{\mu s}}$$
$$\gamma_k = \frac{\frac{\sqrt{\mu s}}{1+\sqrt{\mu s}}\sqrt{\frac{s}{\mu}}}{s} - \left(1 - \mu\sqrt{\frac{s}{\mu}} + \mu\frac{\sqrt{\mu s}}{1+\sqrt{\mu s}}\sqrt{\frac{s}{\mu}}\right) = 0.$$

Thus, AGM-SC converges to the ODE (51) with

$$b(t) = \lim_{s\to 0}\frac{1 - \frac{1-\sqrt{\mu s}}{1+\sqrt{\mu s}}}{\sqrt{s}} = 2\sqrt{\mu}$$
$$c(t) = 0,$$

which recovers AGM-SC ODE.

B.2.3. Difference Matrix and Differential Kernel

The fixed-step first-order scheme (2) with the number of iterations N can be written equivalently as

$$\begin{bmatrix} y_1 - y_0 \\ y_2 - y_1 \\ \vdots \\ y_N - y_{N-1} \end{bmatrix} = -s \begin{bmatrix} h_{0,0} & 0 & \cdots & 0 \\ h_{1,0} & h_{1,1} & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ h_{N-1,0} & h_{N-1,2} & \cdots & h_{N-1,N-1} \end{bmatrix} \begin{bmatrix} \nabla f(y_0) \\ \nabla f(y_1) \\ \vdots \\ \nabla f(y_{N-1}) \end{bmatrix}$$

Here, we call the lower triangular matrix $\mathbf{H} = (h_{ij})$ the *difference matrix* for the algorithm (2).

To derive the limiting ODE of the algorithm (2), we introduce a smooth curve $X : [0, T] \to \mathbb{R}^n$ with the identifications $X(k\sqrt{s}) = y_k$ and $T = N\sqrt{s}$. As a continuous-time analog of the difference matrix (h_{ij}), we intoduce a continuously differentiable function H (possibly depending on s now) defined on $\{(t, \tau) \in \mathbb{R}^2 : 0 < \tau \le t < T\}$ with the identification $H(t_i, \tau_j) = h_{ij}$, where $t_i = i\sqrt{s}$ and $\tau_j = j\sqrt{s}$. Substituting $X(t_i) = y_i$ in (2) yields

$$\frac{X\left(t_{i+1}\right) - X\left(t_i\right)}{\sqrt{s}} = -\left(\tau_{j+1} - \tau_j\right) \sum_{j=0}^{i} H\left(t_i, \tau_j\right) \nabla f\left(X\left(\tau_j\right)\right). \tag{53}$$

Then, we can observe that the right-hand side of (53) is a Riemann sum of the function $\tau \mapsto -H(t_i, \tau)\nabla f(X(\tau))$ over $[0, t_{i+1}]$. Thus, taking the limit $s \to 0$ yields

$$\dot{X}(t) = -\int_0^t H(t, \tau)\nabla f(X(\tau))\, d\tau, \text{ where } H(t, \tau) = \lim_{s\to 0} h_{\frac{t}{\sqrt{s}}, \frac{\tau}{\sqrt{s}}} \tag{54}$$

as the limiting ODE of the fixed-step first-order scheme (2). Therefore, we obtain the ODE (3). Inspired by the observation that the function $H(t, \tau)$ plays a role similar to the *kernel function* in the integral transform, we call it the *differential kernel* (or the *H-kernel*) corresponding to the difference matrix (h_{ij}).

From differential kernels to second-order ODEs. Differentiating both sides of (3) and applying the Leibniz integral rule, we obtain

$$\ddot{X}(t) = -H(t, t)\nabla f(X(t)) - \int_0^t \frac{\partial H(t, \tau)}{\partial t}\nabla f(X(\tau))\, d\tau. \tag{55}$$

If there exists a function $b(t)$ such that

$$\frac{\partial H(t, \tau)}{\partial t} = -b(t)H(t, \tau),$$

then it follows from (3) that the equation (55) is expressed as the following second-order ODE:

$$\ddot{X}(t) + b(t)\dot{X} + H(t, t)\nabla f(X(t)) = 0. \tag{56}$$

From second-order ODEs to differential kernels. For the second-order ODE

$$\ddot{X}(t) + b(t)\dot{X}(t) + (1 + c(t))\nabla f(X(t)) = 0,$$

define the differential kernel $H(t, \tau)$ as

$$H(t, \tau) = (1 + c(\tau))\, e^{-\int_\tau^t b(s)\, ds}. \tag{57}$$

Then, because we have $\frac{\partial H(t,\tau)}{\partial t} = -b(t)H(t, \tau)$ and $H(t, t) = 1 + c(t)$, the ODE (3) with the differential kernel (57) recovers the given second-order ODE.

Recovering the limiting ODE of two-sequence scheme. We first write the two-sequence scheme (49) as the fixed-step first-order scheme. The iterates of the two-sequence scheme (49) satisfy

$$\begin{aligned} y_{k+1} - y_k &= x_{k+1} - y_k + \beta_k\left(x_{k+1} - x_k\right) - s\gamma_k\nabla f\left(y_k\right) \\ &= \beta_k\left(y_k - y_{k-1}\right) + s\beta_k\nabla f\left(y_{k-1}\right) - s\left(1 + \beta_k + \gamma_k\right)\nabla f\left(y_k\right). \end{aligned} \tag{58}$$

Substituting

$$y_{k+1} - y_k = -s\sum_{i=0}^{k} h_{k,i}\nabla f\left(y_i\right),$$

$$y_k - y_{k-1} = -s\sum_{i=0}^{k-1} h_{k-1,i}\nabla f\left(y_i\right)$$

into (58) and comparing the coefficients of each $\nabla f(y_i)$, we obtain

$$h_{k,j} = \begin{cases} 1 + \beta_k + \gamma_k, & \text{if } j = k \\ \beta_k \left(h_{k-1,k-1} - 1 \right), & \text{if } j = k-1 \\ \beta_k h_{k-1,i}, & \text{if } j \leq k-2. \end{cases}$$

Using mathematical induction, it is straightforward to show that

$$h_{ij} = (\beta_j + \gamma_j) \prod_{\nu=j+1}^{i} \beta_\nu + \delta_{ij},$$

where δ_{ij} is the Kronecker delta funciton. For $i > j$,[17] we have

$$h_{i+1,j} - h_{i,j} = (\beta_{i+1} - 1) h_{ij}.$$

Under the identification $H(t_i, \tau_j) = h_{ij}$, we have

$$h_{i+1,j} - h_{i,j} = H(t_{i+1}, \tau_j) - H(t_i, \tau_j) = \frac{\partial H(t_i, \tau_j)}{\partial t} \sqrt{s} + o\left(\sqrt{s}\right).$$

Thus, when the limits (50) exist, taking the limit $s \to 0$ yields

$$\frac{\partial H(t, \tau)}{\partial t} = -b(t) H(t, \tau). \tag{59}$$

Also, because $h_{k+1,k} = \beta_{k+1} + \gamma_k$ and $\lim_{s \to 0} \beta_{t/\sqrt{s}} = 1$ by (50), we have $H(t,t) = 1 + c(t)$ for all $t \in (0, T)$. Therefore, the ODE (56) recovers the limiting ODE (51) of the two-sequence scheme. Moreover, we show that we can explicitly write the differential kernel H as

$$H(t, \tau) = (1 + c(\tau)) \, e^{-\int_\tau^t b(s)\, ds}. \tag{60}$$

By (59), we have

$$\frac{\partial}{\partial s} \log(H(s,t)) = \frac{\partial H(s,\tau)}{\partial s} \frac{1}{H(s,\tau)} = -b(s).$$

Integrating over s, we obtain

$$\log(H(t,\tau)) - \log(H(\tau,\tau)) = -\int_\tau^t b(s)\, ds.$$

Thus, we have

$$H(t,\tau) = H(\tau,\tau) e^{-\int_\tau^t b(s)\, ds} = (1 + c(\tau)) \, e^{-\int_\tau^t b(s)\, ds}.$$

Difference matrix for AGM-C. Because we can write AGM-C as the two-sequence scheme (49) with $\beta_k = \frac{k-1}{k+2}$ and $\gamma_k = 0$, we can rewrite it as the fixed-step first-order scheme (2) with

$$h_{ij} = \prod_{\nu=j}^{i} \frac{\nu - 1}{\nu + 2} + \delta_{ij} = \frac{(j-1)j(j+1)}{i(i+1)(i+2)} + \delta_{ij}.$$

By definition, the differential kernel corresponding to this matrix (h_{ij}) is

$$H(t,\tau) = \lim_{s \to 0} \frac{\left(\frac{\tau}{\sqrt{s}} - 1\right) \frac{\tau}{\sqrt{s}} \left(\frac{\tau}{\sqrt{s}} + 1\right)}{\frac{t}{\sqrt{s}} \left(\frac{t}{\sqrt{s}} + 1\right) \left(\frac{t}{\sqrt{s}} + 2\right)} = \frac{\tau^3}{t^3}. \tag{61}$$

This can be also obtained by substituting $b(t) = 3/t$ and $c(t) = 0$ into (60):

$$H(t,\tau) = e^{-\int_\tau^t \frac{3}{s}\, ds} = e^{-3(\log(t) - \log(\tau))} = \frac{\tau^3}{t^3}.$$

Because we have

$$\frac{\partial H(t,\tau)}{\partial t} = -\frac{3\tau^3}{t^4} = -\frac{3}{t} H(t,\tau),$$

the ODE (55) with the differential kernel (61) recovers AGM-C ODE.

[17] We exclude the case $i = j$ because the difference matrix h_{ij} has singularities at these points due to the Kronecker delta function.

Difference matrix for AGM-SC. Because we can write AGM-SC as the two-sequence scheme (49) with $\beta_k = \frac{1-\sqrt{\mu s}}{1+\sqrt{\mu s}}$ and $\gamma_k = 0$, we can rewrite it as the fixed-step first-order scheme (2) with

$$h_{ij} = \prod_{\nu=j}^{i} \frac{1-\sqrt{\mu s}}{1+\sqrt{\mu s}} + \delta_{ij} = \left(\frac{1-\sqrt{\mu s}}{1+\sqrt{\mu s}}\right)^{i-j+1} + \delta_{ij}.$$

By definition, the differential kernel corresponding to this matrix (h_{ij}) is

$$H(t,\tau) = \lim_{s\to 0} \left(\frac{1-\sqrt{\mu s}}{1+\sqrt{\mu s}}\right)^{\frac{t}{\sqrt{s}}-\frac{\tau}{\sqrt{s}}+1} = \frac{e^{2\sqrt{\mu}\tau}}{e^{2\sqrt{\mu}t}}. \tag{62}$$

This can be also obtained by substituting $b(t) = 2\sqrt{\mu}$ and $c(t) = 0$ into (60):

$$H(t,\tau) = e^{-\int_\tau^t 2\sqrt{\mu}\, ds} = e^{-2\sqrt{\mu}(t-\tau)} = \frac{e^{2\sqrt{\mu}\tau}}{e^{2\sqrt{\mu}t}}.$$

It follows from

$$\frac{\partial H(t,\tau)}{\partial t} = -2\sqrt{\mu}e^{2\sqrt{\mu}(\tau-t)} = -2\sqrt{\mu}H(t,\tau)$$

that the ODE (55) with the differential kernel (62) recovers (AGM-SC ODE).

C. Unified Lagrangian Formulation for Convex and Strongly Convex Objective Functions

C.1. Computing Euler–Lagrange Equation

For the unified Bregman Lagrangian (4), the partial derivatives $\frac{\partial \mathcal{L}}{\partial \dot{X}}\left(X,\dot{X},t\right)$ and $\frac{\partial \mathcal{L}}{\partial X}\left(X,\dot{X},t\right)$ are given by

$$\begin{aligned}
\frac{\partial \mathcal{L}}{\partial \dot{X}}\left(X,\dot{X},t\right) &= e^{\gamma}\left(1+\mu e^{\beta}\right)\left(\nabla h\left(X+e^{-\alpha}\dot{X}\right)-\nabla h(X)\right)\\
\frac{\partial \mathcal{L}}{\partial X}\left(X,\dot{X},t\right) &= e^{\alpha+\gamma}\left(1+\mu e^{\beta}\right)\left(\nabla h\left(X+e^{-\alpha}\dot{X}\right)-\nabla h(X)\right)\\
&\quad - e^{\gamma}\left(1+\mu e^{\beta}\right)\frac{d}{dt}\nabla h(X) - e^{\alpha+\beta+\gamma}\nabla f(X).
\end{aligned}$$

The time derivative of $\frac{\partial \mathcal{L}}{\partial \dot{X}}$ can be computed as

$$\begin{aligned}
\frac{d}{dt}\left\{\frac{\partial \mathcal{L}}{\partial \dot{X}}\left(X,\dot{X},t\right)\right\} &= \left(\dot{\gamma}e^{\gamma}+\mu\left(\dot{\beta}+\dot{\gamma}\right)e^{\beta+\gamma}\right)\left(\nabla h\left(X+e^{-\alpha}\dot{X}\right)-\nabla h(X)\right)\\
&\quad + e^{\gamma}\left(1+\mu e^{\beta}\right)\left(\frac{d}{dt}\nabla h\left(X+e^{-\alpha}\dot{X}\right)-\frac{d}{dt}\nabla h(X)\right).
\end{aligned}$$

Thus, the Euler–Lagrange equation (24) can be written as

$$\begin{aligned}
&e^{\gamma}\left(1+\mu e^{\beta}\right)\frac{d}{dt}\nabla h\left(X+e^{-\alpha}\dot{X}\right)\\
&\quad = \left(e^{\alpha+\gamma}\left(1+\mu e^{\beta}\right)-\dot{\gamma}e^{\gamma}-\mu\left(\dot{\beta}+\dot{\gamma}\right)e^{\beta+\gamma}\right)\left(\nabla h\left(X+e^{-\alpha}\dot{X}\right)-\nabla h(X)\right)-e^{\alpha+\beta+\gamma}\nabla f(X).
\end{aligned}$$

Substituting the ideal scaling condition $\dot{\gamma} = e^{\alpha}$ (23a) into the equation above and then dividing both sides by $e^{\gamma}\left(1+\mu e^{\beta}\right) > 0$, we obtain

$$\frac{d}{dt}\nabla h\left(X+e^{-\alpha}\dot{X}\right) = -\frac{\mu\dot{\beta}e^{\beta}}{1+\mu e^{\beta}}\left(\nabla h\left(X+e^{-\alpha}\dot{X}\right)-\nabla h(X)\right)-\frac{e^{\alpha+\beta}}{1+\mu e^{\beta}}\nabla f(X).$$

Letting $Z = X + e^{-\alpha}\dot{X}$ yields the system of ODEs (5).

C.2. Time Dilation Property of Unified Bregman Lagrangian Flow

We show that the unified Bregman Lagrangian flow (5) is closed under time-dilation, similarly to the first Bregman Lagrangian flow (25) and the second Bregman Lagrangian flow (27) (see Wibisono et al., 2016; Wilson et al., 2021).

Theorem C.1. *Let $\mathbf{T} : I_2 \to I_1$ be a strictly increasing twice-continuously differentiable function, where I_1 and I_2 are intervals in $[0, \infty)$. If (X_1, Z_1) is a solution to the unified Bregman Lagrangian flow (5) on I_1 with the parameters $\alpha_1, \beta_1 : I_1 \to \mathbb{R}$, then the reparametrized curves $X_2(t) := X_1(\mathbf{T}(t))$ and $Z_2(t) := Z_1(\mathbf{T}(t))$ form a solution to the unified Bregman Lagrangian flow on I_2 with the parameters $\alpha_2, \beta_2 : I_2 \to \mathbb{R}$ defined as*

$$\begin{aligned}\alpha_2(t) &= \alpha_1(\mathbf{T}(t)) + \log \dot{\mathbf{T}}(t)\\ \beta_2(t) &= \beta_1(\mathbf{T}(t)).\end{aligned}$$

Proof. Let (X_1, Z_1) be a solution to the system of ODE (5) on I_1 with the parameters $\alpha_1, \beta_1 : I_1 \to \mathbb{R}$. Then, the time derivatives of the curves $X_2(t) = X_1(\mathbf{T}(t))$ and $\nabla h(Z_2(t)) = \nabla h(Z_1(\mathbf{T}(t)))$ can be computed as

$$\begin{aligned}\dot{X}_2(t) &= \dot{\mathbf{T}}(t)\dot{X}_1(\mathbf{T}(t))\\ &= \dot{\mathbf{T}}(t)e^{\alpha_1(\mathbf{T}(t))}(Z_1(\mathbf{T}(t)) - X_1(\mathbf{T}(t))\\ &= \dot{\mathbf{T}}(t)e^{\alpha_1(\mathbf{T}(t))}(Z_2(t) - X_2(t))\\ &= e^{\alpha_2(t)}(Z_2(t) - X_2(t))\end{aligned}$$

and

$$\begin{aligned}\frac{d}{dt}\nabla h(Z_2(t)) &= \dot{\mathbf{T}}(t)\frac{d(\nabla h \circ Z_1)}{dt}(\mathbf{T}(t))\\ &= \dot{\mathbf{T}}(t)\Bigg(\frac{\mu\dot{\beta}_1(\mathbf{T}(t))e^{\beta_1(\mathbf{T}(t))}}{1+\mu e^{\beta_1(\mathbf{T}(t))}}\left(\nabla h(X_1(\mathbf{T}(t)) - \nabla h(Z_1(\mathbf{T}(t)))\right)\\ &\qquad - \frac{e^{\alpha_1(\mathbf{T}(t))+\beta_1(\mathbf{T}(t))}}{1+\mu e^{\beta_1(\mathbf{T}(t))}}\nabla f(X_1(\mathbf{T}(t)))\Bigg)\\ &= \frac{\mu\dot{\beta}_2(t)e^{\beta_2(t)}}{1+\mu e^{\beta_2(t)}}\left(\nabla h(X_2(t)) - \nabla h(Z_2(t))\right) - \frac{e^{\alpha_2(t)+\beta_2(t)}}{1+\mu e^{\beta_2(t)}}\nabla f(X_2(t)).\end{aligned}$$

Thus, (X_2, Z_2) is a solution to the system of ODE (5) on I_2 with the parameters $\alpha_2, \beta_2 : I_2 \to \mathbb{R}$. This completes the proof. □

C.3. Proof of Theorem 3.1

Define the time-dependent Lyapunov function $V : \mathbb{R}^n \times \mathbb{R}^n \times [0, \infty) \to \mathbb{R}$ as

$$V(X, Z, t) = \left(1 + \mu e^{\beta(t)}\right) D_h\left(x^*, Z\right) + e^{\beta(t)}\left(f(X) - f\left(x^*\right)\right). \tag{63}$$

We show that the continuous-time energy function

$$\mathcal{E}(t) = V(X(t), Z(t), t) = \left(1 + \mu e^{\beta(t)}\right) D_h\left(x^*, Z(t)\right) + e^{\beta(t)}\left(f(X(t)) - f\left(x^*\right)\right) \tag{64}$$

is monotonically non-increasing on $[0, \infty)$. Note that

$$\begin{aligned}\frac{d}{dt}D_h\left(x^*, Z\right) &= \frac{d}{dt}\left\{h\left(x^*\right) - h(Z) - \langle\nabla h(Z), x^* - Z\rangle\right\}\\ &= -\left\langle\nabla h(Z), \dot{Z}\right\rangle - \left\langle\frac{d}{dt}\nabla h(Z), x^* - Z\right\rangle + \left\langle\nabla h(Z), \dot{Z}\right\rangle\\ &= -\left\langle\frac{d}{dt}\nabla h(Z), x^* - Z\right\rangle.\end{aligned}$$

Thus, we have

$$\begin{aligned}\frac{d}{dt}\mathcal{E}(t) =& -\left(1+\mu e^{\beta}\right)\left\langle \frac{d}{dt}\nabla h(Z), x^* - Z\right\rangle + \mu\dot{\beta}e^{\beta}D_h\left(x^*, Z\right)\\ &+ \dot{\beta}e^{\beta}\left(f(X) - f\left(x^*\right)\right) + e^{\beta}\left\langle \nabla f(X), \dot{X}\right\rangle\\ =& \left\langle \mu\dot{\beta}e^{\beta}\left(\nabla h(Z) - \nabla h(X)\right) + e^{\alpha+\beta}\nabla f(X), x^* - Z\right\rangle + \mu\dot{\beta}e^{\beta}D_h\left(x^*, Z\right)\\ &+ \dot{\beta}e^{\beta}\left(f(X) - f\left(x^*\right)\right) + e^{\beta}\left\langle \nabla f(X), \dot{X}\right\rangle.\end{aligned}$$

It follows from the Bregman three-point identity,[18] the non-negativity of Bregman divergence, and the μ-uniform convexity of f with respect to h that

$$\begin{aligned}\langle \nabla h(Z) - \nabla h(X), x^* - Z\rangle + D_h\left(x^*, Z\right) &= D_h\left(x^*, X\right) - D_h(Z, X)\\ &\le D_h\left(x^*, X\right)\\ &\le \frac{1}{\mu}D_f\left(x^*, X\right).\end{aligned}$$

Thus, we have

$$\begin{aligned}\frac{d}{dt}\mathcal{E}(t) &\le \dot{\beta}e^{\beta}D_f\left(x^*, X\right) + e^{\alpha+\beta}\left\langle \nabla f(X), x^* - Z\right\rangle\\ &\quad + \dot{\beta}e^{\beta}\left(f(X) - f\left(x^*\right)\right) + e^{\beta}\left\langle \nabla f(X), \dot{X}\right\rangle\\ &= \dot{\beta}e^{\beta}D_f\left(x^*, X\right) + e^{\alpha+\beta}\left\langle \nabla f(X), x^* - X\right\rangle + \dot{\beta}e^{\beta}\left(f(X) - f\left(x^*\right)\right)\\ &= \left(e^{\alpha} - \dot{\beta}\right)e^{\beta}\left\langle \nabla f(X), x^* - X\right\rangle\\ &\le \left(e^{\alpha} - \dot{\beta}\right)e^{\beta}\left(f\left(x^*\right) - f(X)\right)\\ &\le 0,\end{aligned}$$

where the last two inequalities follows from the ideal scaling condition $\dot{\beta}(t) \le e^{\alpha(t)}$ (23b), the convexity of f, and the fact that x^* is a minimizer of f. Writing $\mathcal{E}(t) \le \mathcal{E}(0)$ explicitly completes the proof.

C.4. Second Bregman Lagrangian Flow is Asymptotic Limit of Unified Bregman Lagrangian Flow

For simplicity, we assume that the distance-generating function h is L_h-smooth and μ_h-strongly convex. The μ_h-strong convexity of h implies the L_{h^*}-Lipschitz continuity of ∇h^*, where $L_{h^*} = 1/\mu_h > 0$ (see Rockafellar & Wets, 2009, Proposition 12.60). The following proposition rigorously states that the system (7) is the asymptotic limit of the system (5).

Proposition C.2. *Let $x_0 \in \mathbb{R}^n$ and $T > 0$. Then, for every real number $\epsilon > 0$, there exists a real number $r > 0$ such that*

$$\|X_1(t_0 + T) - X_2(t_0 + T)\| \le \epsilon \quad \forall t_0 > r,$$

where

- *(X_1, Z_1) is the solution to the system (5) with the initial conditions $X_1(t_0) = Z_1(t_0) = x_0$.*
- *(X_2, Z_2) is the solution to the system (7) with the initial conditions $X_2(t_0) = Z_2(t_0) = x_0$.*

Proof. Introducing $W(t) = \nabla h(Z(t))$, the systems (5) and (7) can be equivalently written as

$$\begin{aligned}\dot{X} &= e^{\alpha}(\nabla h^*(W) - X)\\ \dot{W} &= \frac{\mu\dot{\beta}e^{\beta}}{1+\mu e^{\beta}}(\nabla h(X) - W) - \frac{e^{\alpha+\beta}}{1+\mu e^{\beta}}\nabla f(X)\end{aligned} \tag{65}$$

[18] $D_h(x, y) - D_h(x, z) = -\langle \nabla h(y) - \nabla h(z), x - y\rangle - D_h(y, z)$ with $x = x^*$, $y = Z$, $z = X$ (see Wilson et al., 2021).

and

$$\begin{aligned}\dot{X} &= e^{\alpha(\infty)}(\nabla h^*(W) - X)\\ \dot{W} &= \dot{\beta}(\infty)\left(\nabla h(X) - W\right) - \frac{e^{\alpha(\infty)}}{\mu}\nabla f(X),\end{aligned} \tag{66}$$

respectively. Let $W_1(t) = \nabla h(Z_1(t))$ and $W_2(t) = \nabla h(Z_2(t))$. Then, (X_1, W_1) and (X_2, W_2) are the solutions to (65) and (66) respectively. Let $\Delta_X(t) = X_1(t) - X_2(t)$ and $\Delta_W = W_1(t) - W_2(t)$. Then, we have

$$\begin{aligned}\left\|\dot{\Delta}_X(t)\right\| &= \left\|e^{\alpha(\infty)}\left(\nabla h^*(W_1(t)) - \nabla h^*(W_2(t)) - \Delta_X(t)\right) + \left(e^{\alpha(t)} - e^{\alpha(\infty)}\right)\left(\nabla h^*(W_1(t)) - X_1(t)\right)\right\|\\ &\le L_{h^*}e^{\alpha(\infty)}\left\|\Delta_W(t)\right\| + e^{\alpha(\infty)}\left\|\Delta_X(t)\right\| + \left|e^{\alpha(t)} - e^{\alpha(\infty)}\right|\left(\left\|X_1(t)\right\| + \left\|\nabla h^*(0)\right\| + L_{h^*}\left\|W_1(t)\right\|\right)\end{aligned}$$

and

$$\begin{aligned}\left\|\dot{\Delta}_W(t)\right\| &= \Bigg\|\dot{\beta}(\infty)\left(\nabla h(X_1(t)) - \nabla h(X_2(t)) - \Delta_W(t)\right) - \frac{e^{\alpha(\infty)}}{\mu}\left(\nabla f(X_1) - \nabla f(X_2)\right)\\ &\quad + \left(\frac{\mu\dot{\beta}(t)e^{\beta(t)}}{1+\mu e^{\beta(t)}} - \dot{\beta}(\infty)\right)\left(\nabla h(X_1(t)) - W_1(t)\right) - \left(\frac{e^{\alpha(t)+\beta(t)}}{1+\mu e^{\beta(t)}} - \frac{e^{\alpha(\infty)}}{\mu}\right)\nabla f(X_1(t))\Bigg\|\\ &\le \left(L_h\dot{\beta}(\infty) + \frac{L_f e^{\alpha(\infty)}}{\mu}\right)\left\|\Delta_X(t)\right\| + \dot{\beta}(\infty)\left\|\Delta_W(t)\right\|\\ &\quad + \left|\frac{\mu\dot{\beta}(t)e^{\beta(t)}}{1+\mu e^{\beta(t)}} - \dot{\beta}(\infty)\right|\left(\left\|\nabla h(0)\right\| + L_h\left\|X_1(t)\right\| + \left\|W_1(t)\right\|\right)\\ &\quad + \left|\frac{e^{\alpha(t)+\beta(t)}}{1+\mu e^{\beta(t)}} - \frac{e^{\alpha(\infty)}}{\mu}\right|\left(\left\|\nabla f(0)\right\| + L_f\left\|X_1(t)\right\|\right).\end{aligned}$$

We can show that $\|X_1(t)\|$ and $\|Z_1(t)\|$ are bounded above by a constant which is independent of the initial time t_0.[19] Let $U(t) = (\Delta_X(t), \Delta_W(t)) \in \mathbb{R}^{2n}$ and $u(t) = \|U(t)\|$. Let $\delta = \frac{C}{e^{TC}-1}$, where

$$C = L_{h^*}e^{\alpha(\infty)} + e^{\alpha(\infty)} + L_h\dot{\beta}(\infty) + \frac{L_f e^{\alpha(\infty)}}{\mu} + \dot{\beta}(\infty).$$

Then, because $e^{\alpha(t)} \to e^{\alpha(\infty)}$, $\frac{\mu\dot{\beta}(t)e^{\beta(t)}}{1+\mu e^{\beta(t)}} \to \dot{\beta}(\infty)$, and $\frac{e^{\alpha(t)+\beta(t)}}{1+\mu e^{\beta(t)}} \to \frac{e^{\alpha(\infty)}}{\mu}$ as $s \to 0$, we can easily show that:

$$\text{There exists } r \in \mathbb{R} \text{ such that } t \ge t_0 \ge r \Rightarrow u'(t) \le Cu(t) + \delta.$$

Let $v(t) = u(t) + \delta/C$. Then, $\dot{v}(t) \le Cv(t)$ and $v(t_0) = \delta/C$. By Grönwall's inequality, we have

$$\begin{aligned}\left\|X_1(t_0+T) - X_2(t_0+T)\right\| \le u(t_0+T) &= v(t_0+T) - \frac{\delta}{C}\\ &\le v(t_0)\exp\left(\int_{t_0}^{t_0+T} C\right) - \frac{\delta}{C} = v(t_0)e^{TC} - \frac{\delta}{C} = \frac{\delta}{C}\left(e^{TC} - 1\right) = \epsilon\end{aligned}$$

for all $t_0 > r$. This completes the proof. □

C.5. Lyapunov Analysis of Unified Bregman Lagrangian Flow Recovers Lyapunov Analysis of Second Bregman Lagrangian Flow

Because the system (7) is the second Bregman Lagrangian flow (27) with $\alpha_{2\text{nd}}(t) := \alpha(\infty)$ and $\beta_{2\text{nd}}(t) := \dot{\beta}(\infty)t$, its convergence rate can be proven by showing that the following energy function (34) is decreasing:

$$\mathcal{E}_{2\text{nd}}(t) := e^{\dot{\beta}(\infty)t}\left(\mu D_h\left(x^*, Z(t)\right) + f(X(t)) - f\left(x^*\right)\right). \tag{67}$$

[19] This can be proven by bounding $\|X(t) - x^*\|$ and $\|Z(t) - x^*\|$ using the strong convexity of f and the fact that the energy function (83) is non-increasing on $[t_0, \infty)$. We omit the details.

The time derivative of the energy function (64) for the unified Bregman Lagrangian flow can be written as

$$\frac{d}{dt}\mathcal{E}(t) = \frac{d}{dt}1 + \mu e^{\beta} D_h\left(x^*, Z\right) + \left(1 + \mu e^{\beta}\right)\frac{d}{dt}\left\{D_h\left(x^*, Z\right)\right\} + \frac{d}{dt}\left\{e^{\beta}\right\}\left(f(X) - f\left(x^*\right)\right) + e^{\beta}\frac{d}{dt}\left\{f(X) - f\left(x^*\right)\right\}.$$

Because $\frac{d}{dt}\mathcal{E}(t) \leq 0$, we have

$$\begin{aligned} 0 &\geq e^{-\beta(t_0+t)}\frac{d}{dt}\left\{\mathcal{E}(t_0+t)\right\} \\ &= \mu\dot{\beta}(t_0+t) D_h\left(x^*, Z(t_0+t)\right) + \frac{1+\mu e^{\beta(t_0+t)}}{e^{\beta(t_0+t)}}\frac{d}{dt}\left\{D_h\left(x^*, Z(t_0+t)\right)\right\} \\ &\quad + \dot{\beta}(t_0+t)\left(f(X(t_0+t)) - f\left(x^*\right)\right) + \frac{d}{dt}\left\{f(X(t_0+t)) - f\left(x^*\right)\right\} \end{aligned}$$

for all $t > 0$, where $t_0 > 0$ is the initial time of the system. Fix $x_0 = X(t_0) = Z(t_0)$ in $\mathbb{R}^n$. Proposition C.2 shows that as $t_0 \to \infty$, the flow $t \mapsto (X(t_0+t), Z(t_0+t))$ converges to the flow $t \mapsto (X_{\mathrm{2nd}}(t), Z_{\mathrm{2nd}}(t))$ corresponding to the system (7) with $X_{\mathrm{2nd}}(0) = Z_{\mathrm{2nd}}(0) = x_0$ starting at $t = 0$ (because this system is time-invariant, we can shift the initial time). Now, taking the limit $t_0 \to \infty$ in the inequality above yields

$$\begin{aligned} 0 &\geq \mu\dot{\beta}(\infty) D_h\left(x^*, Z(t)\right) + \mu\frac{d}{dt}\left\{D_h\left(x^*, Z(t)\right)\right\} \\ &\quad + \dot{\beta}(\infty)\left(f(X(t)) - f\left(x^*\right)\right) + \frac{d}{dt}\left\{f(X(t)) - f\left(x^*\right)\right\} \\ &= e^{-\beta_{\mathrm{2nd}}(t)}\frac{d}{dt}\mathcal{E}_{\mathrm{2nd}}(t), \end{aligned}$$

where $\mathcal{E}_{\mathrm{2nd}}$ is defined in (67). This completes the proof.

D. Unified ODE Model and Algorithm for Minimizing Convex and Strongly Convex Functions

For convenience, we define the *unified AGM family* as

$$\begin{aligned} y_k &= x_k + \frac{\frac{2\sqrt{s}}{\mathbf{t}_{k+1}}\operatorname{cothc}\left(\frac{\sqrt{\mu}}{2}\mathbf{t}_{k+1}\right) - \mu s}{1-\mu s}\left(z_k - x_k\right) \\ x_{k+1} &= y_k - s\nabla f\left(y_k\right) \\ z_{k+1} &= z_k + \frac{\sqrt{s}\mathbf{t}_{k+1}}{2}\operatorname{tanhc}\left(\frac{\sqrt{\mu}}{2}\mathbf{t}_{k+1}\right)\left(\mu y_k - \mu z_k - \nabla f\left(y_k\right)\right), \end{aligned} \tag{68}$$

where $(\mathbf{t}_k)$ is a strictly increasing sequence in $[0, \infty)$. In particular, it is easy to check that the unified AGM (Algorithm 1) is equivalent to the unified AGM family with the sequence $\mathbf{t}_k := \iota\sqrt{s}k$. This general family will be useful when studying variants of the unified AGM in Appendix D.8.

D.1. Choosing Parameters α and β of Unified Bregman Lagrangian Flow

We first note some properties of the functions α and β that recover AGM-C ODE (or AGM-SC ODE) from the first Bregman Lagrangian flow (or the second Bregman Lagrangian flow, respectively).

The first Bregman Lagrangian flow (25) with $h(x) = \frac{1}{2}\|x\|^2$ can be written as the following ODE:

$$\ddot{X} + \left(-\dot{\alpha} + e^{\alpha}\right)\dot{X} + e^{2\alpha+\beta}\nabla f(X) = 0.$$

The choices $\alpha(t) = \log\frac{2}{t}$ and $\beta(t) = \log\frac{t^2}{4}$, which recover AGM-C ODE, satisfy the ideal scaling condition (23b) with equality and make the coefficient of $\nabla f(X)$ equal to the coefficient of $\ddot{X}$.

The second Bregman Lagrangian flow (27) with $h(x) = \frac{1}{2}\|x\|^2$ can be written as

$$\ddot{X} + \left(-\dot{\alpha} + e^{\alpha} + \dot{\beta}\right)\dot{X} + \frac{e^{2\alpha}}{\mu}\nabla f(X) = 0.$$

The choices $\alpha(t) = \log\sqrt{\mu}$ and $\beta(t) = \log\left(\sqrt{\mu}t\right)$, which recover AGM-SC ODE, satisfy the ideal scaling condition (23b) with equality and make the coefficient of $\nabla f(X)$ equal to the coefficient of $\ddot{X}$.

Inspired by these facts, in the unified Bregman Lagrangian, we construct functions $\alpha(t)$ and $\beta(t)$ so that the ideal scaling condition (23b) holds with equality and that the coefficient of $\nabla f(X)$ is equal to the coefficient of $\ddot{X}$. The unified Bregman Lagrangian flow (5) with $h(x) = \frac{1}{2}\|x\|^2$ can be equivalently written as

$$\ddot{X} + \left(-\dot{\alpha} + e^{\alpha} + \frac{\mu\dot{\beta}e^{\beta}}{1+\mu e^{\beta}}\right)\dot{X} + \frac{e^{2\alpha+\beta}}{1+\mu e^{\beta}}\nabla f(X) = 0.$$

Now, the conditions aforementioned can be written as the following system of ODEs:

$$\begin{aligned}\dot{\beta} &= e^{\alpha}\\ e^{2\alpha+\beta} &= 1 + \mu e^{\beta}.\end{aligned}$$

We solve this system. Let $A(t) = e^{\beta(t)} > 0$. Then, we have $\dot{A} = \dot{\beta}e^{\beta} = e^{\alpha+\beta} > 0$. Because $(\dot{A})^2 = e^{2\alpha+\beta}e^{\beta} = A(1+\mu A)$, we have $\dot{A} = \sqrt{A(1+\mu A)}$. Solving this differential equation with the initial condition $A(0) = 0$ yields $A(t) = \frac{t^2}{4}\operatorname{sinhc}^2(\frac{\sqrt{\mu}}{2}t)$. In this case, we have $\beta(t) = \log(\frac{t^2}{4}\operatorname{sinhc}^2(\frac{\sqrt{\mu}}{2}t))$ and $\alpha(t) = \log(\dot{\beta}(t)) = \log(\frac{2}{t}\operatorname{cothc}(\frac{\sqrt{\mu}}{2}t))$.

D.2. Equivalent Forms of Unified AGM ODE

In this section, we provide the three equivalent forms of the unified AGM ODE (8). We assume $\mu > 0$ for the sake of simplicity. The unified Bregman Lagrangian flow (5) with $h(x) = \frac{1}{2}\|x\|^2$, $\alpha(t) = \log(\frac{2}{t}\operatorname{cothc}(\frac{\sqrt{\mu}}{2}t))$, and $\beta(t) = \log(\frac{t^2}{4}\operatorname{sinhc}^2(\frac{\sqrt{\mu}}{2}t))$ can be written as the following system of ODEs, which we call the *unified AGM system*:

$$\begin{aligned}\dot{X} &= \frac{2}{t}\operatorname{cothc}\left(\frac{\sqrt{\mu}}{2}t\right)(Z - X)\\ \dot{Z} &= \frac{t}{2}\operatorname{tanhc}\left(\frac{\sqrt{\mu}}{2}t\right)(\mu X - \mu Z - \nabla f(X)).\end{aligned} \tag{69}$$

In what follows, we show that this system is equivalent to the unified AGM ODE (8).

Second-order ODE form of the unified AGM system. When $\mu > 0$, we can write the unified AGM system (69) as

$$\begin{aligned}\dot{X} &= \sqrt{\mu}\coth\left(\frac{\sqrt{\mu}}{2}t\right)(Z - X)\\ \dot{Z} &= \frac{1}{\sqrt{\mu}}\tanh\left(\frac{\sqrt{\mu}}{2}t\right)(\mu X - \mu Z - \nabla f(X)).\end{aligned}$$

Substituting $Z = X + \frac{1}{\sqrt{\mu}}\tanh(\frac{\sqrt{\mu}}{2}t)\dot{X}$ into $\dot{Z} = \frac{1}{\sqrt{\mu}}\tanh(\frac{\sqrt{\mu}}{2}t)(\mu X - \mu Z - \nabla f(X))$, we obtain

$$\begin{aligned}&\frac{1}{\sqrt{\mu}}\tanh\left(\frac{\sqrt{\mu}}{2}t\right)\ddot{X} + \left(1 + \frac{1}{2}\operatorname{sech}^2\left(\frac{\sqrt{\mu}}{2}t\right)\right)\\ &= \frac{1}{\sqrt{\mu}}\tanh\left(\frac{\sqrt{\mu}}{2}t\right)(\mu X - \mu Z - \nabla f(X))\\ &= -\sqrt{\mu}\tanh\left(\frac{\sqrt{\mu}}{2}t\right)(Z - X) - \frac{1}{\sqrt{\mu}}\tanh\left(\frac{\sqrt{\mu}}{2}t\right)\nabla f(X)\\ &= -\tanh^2\left(\frac{\sqrt{\mu}}{2}t\right)\dot{X} - \frac{1}{\sqrt{\mu}}\tanh\left(\frac{\sqrt{\mu}}{2}t\right)\nabla f(X).\end{aligned}$$

Multiplying by $\sqrt{\mu}\coth(\frac{\sqrt{\mu}}{2}t)$ and rearranging the terms, we have

$$\ddot{X} + \left(\sqrt{\mu}\tanh\left(\frac{\sqrt{\mu}}{2}t\right) + \sqrt{\mu}\coth\left(\frac{\sqrt{\mu}}{2}t\right) + \frac{\sqrt{\mu}}{2}\operatorname{sech}\left(\frac{\sqrt{\mu}}{2}t\right)\operatorname{csch}\left(\frac{\sqrt{\mu}}{2}t\right)\right)\dot{X} + \nabla f(X) = 0.$$

Using the identity $\tanh(x) - \coth(x) + \operatorname{sech}(x)\operatorname{csch}(x) = 0$, we can equivalently write this ODE as

$$\ddot{X} + \left(\frac{\sqrt{\mu}}{2}\tanh\left(\frac{\sqrt{\mu}}{2}t\right) + \frac{3\sqrt{\mu}}{2}\coth\left(\frac{\sqrt{\mu}}{2}t\right)\right)\dot{X} + \nabla f(X) = 0,$$

which is the unified AGM ODE (8).

Differential kernel of the unified AGM ODE. Substituting $b(t) = \frac{\sqrt{\mu}}{2}\tanh(\frac{\sqrt{\mu}}{2}t) + \frac{3\sqrt{\mu}}{2}\coth(\frac{\sqrt{\mu}}{2}t)$ and $c(t) = 0$ into (60), we yield the differential kernel of the unified AGM ODE as

$$\begin{aligned} H(t,\tau) &= e^{-\int_\tau^t \left(\frac{\sqrt{\mu}}{2}\tanh\left(\frac{\sqrt{\mu}}{2}s\right) + \frac{3\sqrt{\mu}}{2}\coth\left(\frac{\sqrt{\mu}}{2}s\right)\right)ds} \\ &= e^{-\left[3\log\left(\sinh\left(\frac{\sqrt{\mu}}{2}s\right)\right) + \log\left(\cosh\left(\frac{\sqrt{\mu}}{2}s\right)\right)\right]_\tau^t} \\ &= \frac{\sinh^3\left(\frac{\sqrt{\mu}}{2}\tau\right)\cosh\left(\frac{\sqrt{\mu}}{2}\tau\right)}{\sinh^3\left(\frac{\sqrt{\mu}}{2}t\right)\cosh\left(\frac{\sqrt{\mu}}{2}t\right)}. \end{aligned}$$

D.3. Existence and Uniqueness of Solution to Unified AGM ODE

In order to prove the existence and uniqueness of a solution to the unified AGM system (69) (which is equivalent to the unified AGM ODE), we prove a stronger result, that the unified Bregman Lagrangian flow (5) with $\alpha(t) = \log(\frac{2}{t}\operatorname{cothc}(\frac{\sqrt{\mu}}{2}t))$ and $\beta(t) = \log(\frac{t^2}{4}\operatorname{sinhc}^2(\frac{\sqrt{\mu}}{2}t))$:

$$\begin{aligned} \dot{X} &= \frac{2}{t}\operatorname{cothc}\left(\frac{\sqrt{\mu}}{2}t\right)(Z - X) \\ \frac{d}{dt}\nabla h(Z) &= \frac{t}{2}\operatorname{tanhc}\left(\frac{\sqrt{\mu}}{2}t\right)(\mu\nabla h(X) - \mu\nabla h(Z) - \nabla f(X)) \end{aligned} \tag{70}$$

with the initial conditions $X(0) = Z(0) = x_0$ has a unique global solution (X, Z) in $C^1([0,\infty), \mathbb{R}^n \times \mathbb{R}^n)$. Following Krichene et al. (2015), we assume that ∇f is L_f-Lipschitz continuous, ∇h is L_h-Lipschitz continuous, and that h is μ_h-strongly convex. The μ_h-strong convexity of h implies the L_{h^*}-Lipschitz continuity of ∇h^*, where $L_{h^*} = \frac{1}{\mu_h} > 0$ (see Rockafellar & Wets, 2009, Proposition 12.60).

D.3.1. Proof of Existence

Fix $t_1 > 0$. We show the existence of solution to the system (70) on $[0, t_1]$. To remove the singularity of the system (70) at $t = 0$, fix $\delta > 0$, and consider the following system of ODEs:

$$\begin{aligned} \dot{X} &= \frac{2}{\max\{\delta, t\}}\operatorname{cothc}\left(\frac{\sqrt{\mu}}{2}\max\{\delta, t\}\right)(Z - X) \\ \frac{d}{dt}\nabla h(Z) &= \frac{t}{2}\operatorname{tanhc}\left(\frac{\sqrt{\mu}}{2}t\right)(\mu\nabla h(X) - \mu\nabla h(Z) - \nabla f(X)) \end{aligned} \tag{71}$$

with $X(0) = Z(0) = x_0$. Denote the image of Z under the mirror map ∇h as $W(t) = \nabla h(Z(t))$. Denote the convex conjugate of h by $h^* : \mathbb{R}^n \to \mathbb{R}$. Then, ∇h and ∇h^* are inverses of each other (see Rockafellar & Wets, 2009, Section 11). Now, we can equivalently write the system (71) as

$$\dot{X} = \frac{2}{\max\{\delta, t\}}\operatorname{cothc}\left(\frac{\sqrt{\mu}}{2}\max\{\delta, t\}\right)(\nabla h^*(W) - X) \tag{72a}$$

$$\dot{W} = \frac{t}{2}\operatorname{tanhc}\left(\frac{\sqrt{\mu}}{2}t\right)\left(\mu\nabla h(X) - \mu W - \nabla f(X)\right) \tag{72b}$$

with the $X(0) = x_0$ and $W(0) = w_0 := \nabla h(x_0)$. By the Cauchy-Lipschitz theorem (Teschl, 2012, Theorem 25), the system of ODEs (72) has a unique solution (X_δ, W_δ) in $C^1([0, t_1], \mathbb{R}^n \times \mathbb{R}^n)$. If we prove the following lemma, then one can prove the existence of solution to the ODE system (71) following the argument in (Krichene et al., 2015, Section 3.2).

Lemma D.1. *Define a constant T as*

$$T = \min\left\{\sqrt{\frac{2}{\mu}}, \frac{1}{2}\sqrt{\frac{1}{K_2 K_3}}\right\},$$

where K_2 and K_3 are constants defined in (74). Then, the family of solutions $((X_\delta, Z_\delta)|_{[0,T]})_{\delta\in(0,T]}$ is equi-Lipschitz-continuous and uniformly bounded.

We now prove this lemma. We follow the argument of Krichene et al. (2015) and omit the detailed calculations that can be found in (Krichene et al., 2015, Appendix 2). Fix δ. For $t > 0$, define

$$\begin{aligned}
A_\delta(t) &:= \sup_{u\in[0,t]} \frac{\left\|\dot{W}_\delta(u)\right\|}{u} \\
B_\delta(t) &:= \sup_{u\in[0,t]} \frac{\|X_\delta(u) - x_0\|}{u} \\
C_\delta(t) &:= \sup_{u\in[0,t]} \left\|\dot{X}_\delta(u)\right\|.
\end{aligned}$$

Then, these quantities are finite. We first prove the following inequalities, which correspond to (Krichene et al., 2015, Lemma 3):

$$A_\delta(t) \le \mu\|w_0\| + \mu\|\nabla h(x_0)\| + \|\nabla f(x_0)\| + (\mu L_h + L_f)\, t B_\delta(t) \tag{73a}$$

$$B_\delta(t) \le \frac{L_{h^*} t}{3}\operatorname{cothc}\left(\frac{\sqrt{\mu}}{2}T\right) A_\delta(t) \tag{73b}$$

$$C_\delta(t) \le \operatorname{cothc}\left(\frac{\sqrt{\mu}}{2}T\right)\left(L_{h^*} T A_\delta(t) + 2B_\delta(t)\right). \tag{73c}$$

Proof of (73a). Using A_δ and B_δ, we can bound $\|W_\delta(t) - w_0\|$ and $\|X_\delta(t) - x_0\|$ as

$$\begin{aligned}
\|W_\delta(t) - w_0\| &\le \frac{t^2}{2} A_\delta(t) \\
\|X_\delta(t) - x_0\| &\le t B_\delta(t).
\end{aligned}$$

From (72b), we have

$$\begin{aligned}
2\frac{\left\|\dot{W}_\delta(t)\right\|}{t} &= \operatorname{tanhc}\left(\frac{\sqrt{\mu}}{2}t\right)\|\mu\nabla h(X_\delta) - \mu W_\delta - \nabla f(X_\delta)\| \\
&\le \|\mu\nabla h(X_\delta) - \mu W_\delta - \nabla f(X_\delta)\| \\
&\le \mu\|W_\delta\| + \mu\|\nabla h(X_\delta)\| + \|\nabla f(X_\delta)\| \\
&\le \mu\|w_0\| + \frac{\mu t^2}{2} A_\delta(t) + \mu\|\nabla h(x_0)\| + \mu L_h t B_\delta(t) + \|\nabla f(x_0)\| + L_f t B_\delta(t).
\end{aligned}$$

Thus,

$$\begin{aligned}
2A_\delta(t) \le{}& \mu\|w_0\| + \mu\|\nabla h(x_0)\| + \|\nabla f(x_0)\| \\
&+ \frac{\mu t^2}{2} A_\delta(t) + (\mu L_h + L_f)\, t B_\delta(t).
\end{aligned}$$

Because $T \le \sqrt{2/\mu}$, we obtain the inequality (73a).

Proof of (73b). To bound the function $B_\delta(t) = \sup_{u\in[0,t]} \frac{\|X_\delta(u)-x_0\|}{u}$, we first compute an upper bound of $\|X_\delta(t) - x_0\|$ in the case $0 \le t \le \delta$ and the case $t \ge \delta$ separately. First, consider the case $t \in [0, \delta]$. By (72a), we have

$$\dot{X}_\delta + \frac{2}{\delta}\operatorname{cothc}\left(\frac{\sqrt{\mu}}{2}\delta\right)(X_\delta - x_0) = \frac{2}{\delta}\operatorname{cothc}\left(\frac{\sqrt{\mu}}{2}\delta\right)\left(\nabla h^*(W_\delta - \nabla h^*(w_0)\right).$$

Multiplying $e^{\frac{2}{\delta}\operatorname{cothc}\left(\frac{\sqrt{\mu}}{2}\delta\right)t}$, we obtain

$$e^{\frac{2}{\delta}\operatorname{cothc}\left(\frac{\sqrt{\mu}}{2}\delta\right)t}\left[\dot{X}_\delta + \frac{2}{\delta}\operatorname{cothc}\left(\frac{\sqrt{\mu}}{2}\delta\right)(X_\delta - x_0)\right] = \frac{2}{\delta}\operatorname{cothc}\left(\frac{\sqrt{\mu}}{2}\delta\right)e^{\frac{2}{\delta}\operatorname{cothc}\left(\frac{\sqrt{\mu}}{2}\delta\right)t}\left(\nabla h^*(W_\delta) - \nabla h^*(w_0)\right).$$

This equality can be written as

$$\frac{d}{dt}\left((X_\delta(t) - x_0)\, e^{\frac{2}{\delta}\operatorname{cothc}\left(\frac{\sqrt{\mu}}{2}\delta\right)t}\right) = \frac{2}{\delta}\operatorname{cothc}\left(\frac{\sqrt{\mu}}{2}\delta\right)e^{\frac{2}{\delta}\operatorname{cothc}\left(\frac{\sqrt{\mu}}{2}\delta\right)t}\left(\nabla h^*(W_\delta(t)) - \nabla h^*(w_0)\right).$$

Integrating both sides yields

$$(X_\delta(t) - x_0)\, e^{\frac{2}{\delta}\operatorname{cothc}\left(\frac{\sqrt{\mu}}{2}\delta\right)t} = \frac{2}{\delta}\operatorname{cothc}\left(\frac{\sqrt{\mu}}{2}\delta\right)\int_0^t \left[e^{\frac{2}{\delta}\operatorname{cothc}\left(\frac{\sqrt{\mu}}{2}\delta\right)s}\left(\nabla h^*(W_\delta(s)) - \nabla h^*(w_0)\right)\right] ds.$$

Taking norms, we have

$$\begin{aligned}
\|X_\delta(t) - x_0\| &\le \frac{2}{\delta}\operatorname{cothc}\left(\frac{\sqrt{\mu}}{2}\delta\right)\int_0^t \|\nabla h^*(W_\delta(s)) - \nabla h^*(w_0)\|\, ds \\
&\le \frac{2L_{h^*}}{\delta}\operatorname{cothc}\left(\frac{\sqrt{\mu}}{2}\delta\right)\int_0^t \|W_\delta(s) - w_0\|\, ds \\
&\le \frac{2L_{h^*}}{\delta}\operatorname{cothc}\left(\frac{\sqrt{\mu}}{2}\delta\right)\int_0^t \frac{s^2}{2}A_\delta(t)\, ds \\
&= \frac{2L_{h^*}}{\delta}\operatorname{cothc}\left(\frac{\sqrt{\mu}}{2}\delta\right)A_\delta(t)\frac{t^3}{6} \\
&\le \frac{2L_{h^*}}{t}\operatorname{cothc}\left(\frac{\sqrt{\mu}}{2}\delta\right)A_\delta(t)\frac{t^3}{6} \\
&= \frac{L_{h^*}t^2}{3}\operatorname{cothc}\left(\frac{\sqrt{\mu}}{2}\delta\right)A_\delta(t).
\end{aligned}$$

So far, we provide an upper bound of $\|X_\delta(t) - x_0\|$ in the case $0 \le t \le \delta$. We now consider the case $t \ge \delta$. By (72a), we have

$$\dot{X}_\delta + \frac{2}{t}\operatorname{cothc}\left(\frac{\sqrt{\mu}}{2}t\right)(X_\delta - x_0) = \frac{2}{t}\operatorname{cothc}\left(\frac{\sqrt{\mu}}{2}t\right)\left(\nabla h^*(W_\delta) - \nabla h^*(w_0)\right).$$

Multiplying $\frac{t^2}{4}\operatorname{sinhc}^2(\frac{\sqrt{\mu}}{2}t)$ to both sides, we obtain

$$\begin{aligned}
&\frac{t^2}{4}\operatorname{sinhc}^2\left(\frac{\sqrt{\mu}}{2}t\right)\dot{X}_\delta + \frac{t}{2}\operatorname{sinhc}\left(\frac{\sqrt{\mu}}{2}t\right)\cosh\left(\frac{\sqrt{\mu}}{2}t\right)(X_\delta - x_0) \\
&\qquad = \frac{t}{2}\operatorname{sinhc}\left(\frac{\sqrt{\mu}}{2}t\right)\cosh\left(\frac{\sqrt{\mu}}{2}t\right)\left(\nabla h^*(W_\delta) - \nabla h^*(w_0)\right).
\end{aligned}$$

This equality can be written as

$$\frac{d}{dt}\left(\frac{t^2}{4}\operatorname{sinhc}^2\left(\frac{\sqrt{\mu}}{2}t\right)(X_\delta(t) - x_0)\right) = \frac{t}{2}\operatorname{sinhc}\left(\frac{\sqrt{\mu}}{2}t\right)\cosh\left(\frac{\sqrt{\mu}}{2}t\right)\left(\nabla h^*(W_\delta(t)) - \nabla h^*(w_0)\right).$$

Integrating both sides, we obtain

$$\frac{t^2}{4}\operatorname{sinhc}^2\left(\frac{\sqrt{\mu}}{2}t\right)(X_\delta(t) - x_0) = \int_0^t \left(\frac{s}{2}\operatorname{sinhc}\left(\frac{\sqrt{\mu}}{2}s\right)\cosh\left(\frac{\sqrt{\mu}}{2}s\right)\left(\nabla h^*(W_\delta(s)) - \nabla h^*(w_0)\right)\right) ds.$$

Taking norms, we have the following upper bound on $\|X_\delta(t) - x_0\|$:

$$\begin{aligned}
\|X_\delta(t) - x_0\| &\le \frac{2}{t}\operatorname{cothc}\left(\frac{\sqrt{\mu}}{2}t\right)\int_0^t \|\nabla h^*\left(W_\delta(s)\right) - \nabla h^*\left(w_0\right)\|\, ds. \\
&\le \frac{2L_{h^*}}{t}\operatorname{cothc}\left(\frac{\sqrt{\mu}}{2}t\right)\int_0^t \|W_\delta(s) - w_0\|\, ds. \\
&\le \frac{2L_{h^*}}{t}\operatorname{cothc}\left(\frac{\sqrt{\mu}}{2}t\right)\int_0^t \frac{s^2}{2}A_\delta(t)\, ds \\
&= \frac{2L_{h^*}}{t}\operatorname{cothc}\left(\frac{\sqrt{\mu}}{2}t\right)A_\delta(t)\frac{t^3}{6} \\
&= \frac{L_{h^*}t^2}{3}\operatorname{cothc}\left(\frac{\sqrt{\mu}}{2}t\right)A_\delta(t).
\end{aligned}$$

Combining both cases $0 \le t \le \delta$ and $t \ge \delta$, we have

$$\|X_\delta(t) - x_0\| \le \frac{L_{h^*}t^2}{3}\operatorname{cothc}\left(\frac{\sqrt{\mu}}{2}T\right)A_\delta(t)$$

for all $t \ge 0$. Dividing by t and taking the supremum, we obtain

$$B_\delta(t) \le \frac{L_{h^*}t}{3}\operatorname{cothc}\left(\frac{\sqrt{\mu}}{2}T\right)A_\delta(t).$$

Proof of (73c). By (72a), we have

$$\begin{aligned}
\left\|\dot{X}\right\| &= \frac{2}{\max\{\delta,t\}}\operatorname{cothc}\left(\frac{\sqrt{\mu}}{2}\max\{\delta,t\}\right)\|\nabla h^*\left(W_\delta(t)\right) - X_\delta(t)\| \\
&\le \frac{2}{\max\{\delta,t\}}\operatorname{cothc}\left(\frac{\sqrt{\mu}}{2}\max\{\delta,t\}\right)\left(\|\nabla h^*\left(W_\delta(t)\right) - \nabla h^*\left(z_0\right)\| + \|X_\delta(t) - x_0\|\right) \\
&\le \frac{2}{\max\{\delta,t\}}\operatorname{cothc}\left(\frac{\sqrt{\mu}}{2}\max\{\delta,t\}\right)\left(\frac{t^2}{2}L_{h^*}A_\delta(t) + tB_\delta(t)\right) \\
&\le \operatorname{cothc}\left(\frac{\sqrt{\mu}}{2}T\right)\frac{2}{t}\left(\frac{t^2}{2}L_{h^*}A_\delta(t) + tB_\delta(t)\right) \\
&\le \operatorname{cothc}\left(\frac{\sqrt{\mu}}{2}T\right)\left(L_{h^*}TA_\delta(t) + 2B_\delta(t)\right).
\end{aligned}$$

Complete the proof of Lemma D.1. Define five positive constants $K_1, \ldots, K_5$ as

$$\begin{aligned}
K_1 &:= \mu\|w_0\| + \mu\|\nabla h\left(x_0\right)\| + \|\nabla f\left(x_0\right)\| \\
K_2 &:= \mu L_h + L_f \\
K_3 &:= \frac{2L_{h^*}}{3} \\
K_4 &:= 2L_{h^*} \\
K_5 &:= 4.
\end{aligned} \tag{74}$$

Because $T \le \frac{2}{\sqrt{\mu}}$, we have $\operatorname{cothc}(\frac{\sqrt{\mu}}{2}T) \le \operatorname{cothc}(1) \le 2$. Thus, the inequalities (73) imply

$$A_\delta(t) \le K_1 + K_2TB_\delta(t) \tag{75a}$$
$$B_\delta(t) \le K_3TA_\delta(t) \tag{75b}$$
$$C_\delta(t) \le K_4TA_\delta(t) + K_5B_\delta(t). \tag{75c}$$

Combining (75a) and (75b), we have

$$\left(\frac{1}{K_3 T} - K_2 T\right) B_\delta(t) \leq K_1.$$

Because $T \mapsto \frac{1}{K_3 T} - K_2 T$ is a positive non-increasing funtion on $[0, \frac{1}{2}\sqrt{\frac{1}{K_2 K_3}}]$ and $T \leq \frac{1}{2}\sqrt{\frac{1}{K_2 K_3}}$, we have

$$B_\delta(T) \leq \left(\frac{1}{K_3 \cdot \frac{1}{2}\sqrt{\frac{1}{K_2 K_3}}} - K_2 \cdot \frac{1}{2}\sqrt{\frac{1}{K_2 K_3}}\right)^{-1} K_1 = \frac{2}{3} K_1 \sqrt{\frac{K_3}{K_2}}. \tag{76}$$

The inequalities (75a), (76), and $T \leq \frac{1}{2}\sqrt{\frac{1}{K_2 K_3}}$ imply

$$A_\delta(T) \leq K_1 + K_2 T B_\delta(T) \leq K_1 + K_2 \left(\frac{1}{2}\sqrt{\frac{1}{K_2 K_3}}\right)\left(\frac{2}{3} K_1 \sqrt{\frac{K_3}{K_2}}\right). \tag{77}$$

The inequalities (75a), (76), (77), and $T \leq \frac{1}{2}\sqrt{\frac{1}{K_2 K_3}}$ imply

$$\begin{aligned} C_\delta(T) &\leq K_4 T A_\delta(T) + K_5 B_\delta(T) \\ &\leq K_4 \left(\frac{1}{2}\sqrt{\frac{1}{K_2 K_3}}\right)\left(K_1 + K_2 \left(\frac{1}{2}\sqrt{\frac{1}{K_2 K_3}}\right)\left(\frac{2}{3} K_1 \sqrt{\frac{K_3}{K_2}}\right)\right) + K_5 \left(\frac{2}{3} K_1 \sqrt{\frac{K_3}{K_2}}\right). \end{aligned} \tag{78}$$

Therefore, $\|\dot{W}\|$ and $\|\dot{X}\|$ are bounded uniformly in δ because

$$\begin{aligned} \left\|\dot{W}_\delta(t)\right\| &\leq T A_\delta(T) \\ \left\|\dot{X}_\delta(t)\right\| &\leq C_\delta(T) \end{aligned}$$

for all $t \in [0, T]$. This implies that the family of solutions $((X_\delta, Z_\delta)|_{[0,T]})_{\delta \in (0,T]}$ is equi-Lipschitz-continuous and uniformly bounded.

D.3.2. Proof of Uniqueness

We follow the argument in (Krichene et al., 2015, Appendix 3) and omit the detailed calculations that can be found in (Krichene et al., 2015, Appendix 3). Because we only need to prove the uniqueness of solution near $t = 0$, we assume $t < T$ for some $T > 0$. Let (X, W) and $\left(\bar{X}, \bar{W}\right)$ be solutions to the following system of ODEs, which is equivalent to (70):

$$\begin{aligned} \dot{X} &= \frac{2}{t} \operatorname{cothc}\left(\frac{\sqrt{\mu}}{2} t\right)(\nabla h^*(W) - X) \\ \dot{W} &= \frac{t}{2} \operatorname{tanhc}\left(\frac{\sqrt{\mu}}{2} t\right)(\mu \nabla h(X) - \mu W - \nabla f(X)). \end{aligned}$$

Let $\Delta_W = W - \bar{W}$ and $\Delta_X = X - \bar{X}$. Then, we have

$$\begin{aligned} \dot{\Delta}_W &= \frac{t}{2} \operatorname{tanhc}\left(\frac{\sqrt{\mu}}{2} t\right)\left(\mu \nabla h(X) - \mu W - \nabla f(X) - \mu \nabla h\left(\bar{X}\right) + \mu \bar{W} + \nabla f\left(\bar{X}\right)\right) \\ \dot{\Delta}_X &= \frac{2}{t} \operatorname{cothc}\left(\frac{\sqrt{\mu}}{2} t\right)\left(\nabla h^*(W) - \nabla h^*\left(\bar{W}\right) - \Delta_X\right) \end{aligned}$$

with $\Delta_X(0) = \Delta_W(0) = 0$. Define

$$A(t) := \sup_{[0,t]} \frac{\left\|\dot{\Delta}_W(u)\right\|}{u}$$

$$B(t) := \sup_{[0,t]} \|\Delta_X\|.$$

Then, $B(t)$ and $C(t)$ are finite because Δ_X and Δ_W are continuous. First, we compute an upper bound of $A(t)$. We have

$$\begin{aligned}
\left\|\dot{\Delta}_W(t)\right\| &= \frac{t}{2}\operatorname{tanhc}\left(\frac{\sqrt{\mu}}{2}t\right)\left\|\mu\nabla h(X) - \mu W - \nabla f(X) - \mu\nabla h\left(\bar{X}\right) + \mu\bar{W} + \nabla f\left(\bar{X}\right)\right\| \\
&\leq \frac{t}{2}\operatorname{tanhc}\left(\frac{\sqrt{\mu}}{2}t\right)\left(\mu\left\|\nabla h(X) - \nabla h\left(\bar{X}\right)\right\| + \mu\left\|W - \bar{W}\right\| + \left\|\nabla f(X) - \nabla f\left(\bar{X}\right)\right\|\right) \\
&\leq \frac{t}{2}\operatorname{tanhc}\left(\frac{\sqrt{\mu}}{2}t\right)\left(\left(\mu L_h + L_f\right)\|\Delta_X\| + \mu\|\Delta_W\|\right) \\
&\leq \frac{t}{2}\operatorname{tanhc}\left(\frac{\sqrt{\mu}}{2}t\right)\left(\left(\mu L_h + L_f\right)B(t) + \frac{\mu t^2}{2}A(t)\right),
\end{aligned} \tag{79}$$

where we used $\|\Delta_W(t)\| \leq \|\int_0^t \dot{\Delta}_W(s)\,ds\| \leq \int_0^t sA(s)\,ds \leq \int_0^t sA(t)\,ds = \frac{t^2}{2}A(t)$ for the last inequality. Dividing both sides of (79) by t and then taking the supremum, we obtain

$$A(t) \leq \frac{1}{2}\operatorname{tanhc}\left(\frac{\sqrt{\mu}}{2}t\right)\left(\left(\mu L_h + L_f\right)B(t) + \frac{\mu t^2}{2}A(t)\right). \tag{80}$$

Nest, we compute an upper bound of $B(t)$. We have

$$\dot{\Delta}_X + \frac{2}{t}\operatorname{cothc}\left(\frac{\sqrt{\mu}}{2}t\right)\Delta_X = \frac{2}{t}\operatorname{cothc}\left(\frac{\sqrt{\mu}}{2}t\right)\left(\nabla h^*(W) - \nabla h^*\left(\bar{W}\right)\right).$$

Multiplying both sides by $\frac{t^2}{4}\operatorname{sinhc}^2\left(\frac{\sqrt{\mu}}{2}t\right)$, we have

$$\begin{aligned}
&\frac{t^2}{4}\operatorname{sinhc}^2\left(\frac{\sqrt{\mu}}{2}t\right)\dot{\Delta}_X + \frac{t}{2}\operatorname{sinhc}\left(\frac{\sqrt{\mu}}{2}t\right)\cosh\left(\frac{\sqrt{\mu}}{2}t\right)\Delta_X \\
&\qquad = \frac{t}{2}\operatorname{sinhc}\left(\frac{\sqrt{\mu}}{2}t\right)\cosh\left(\frac{\sqrt{\mu}}{2}t\right)\left(\nabla h^*(W) - \nabla h^*\left(\bar{W}\right)\right).
\end{aligned}$$

This equality can be written as

$$\frac{d}{dt}\left(\frac{t^2}{4}\operatorname{sinhc}^2\left(\frac{\sqrt{\mu}}{2}t\right)\Delta_X\right) = \frac{t}{2}\operatorname{sinhc}\left(\frac{\sqrt{\mu}}{2}t\right)\cosh\left(\frac{\sqrt{\mu}}{2}t\right)\left(\nabla h^*(W) - \nabla h^*\left(\bar{W}\right)\right).$$

Integrating both sides, we obtain

$$\frac{t^2}{4}\operatorname{sinhc}^2\left(\frac{\sqrt{\mu}}{2}t\right)\Delta_X = \int_0^t\left[\frac{s}{2}\operatorname{sinhc}\left(\frac{\sqrt{\mu}}{2}s\right)\cosh\left(\frac{\sqrt{\mu}}{2}s\right)\left(\nabla h^*(W(s)) - \nabla h^*\left(\bar{W}(s)\right)\right)\right]ds.$$

Taking norms, we have

$$\begin{aligned}
\|\Delta_X(t)\| &\leq \frac{2}{t}\operatorname{cothc}\left(\frac{\sqrt{\mu}}{2}t\right)\int_0^t\left\|\nabla h^*(W(s)) - \nabla h^*\left(\bar{W}(s)\right)\right\|ds. \\
&\leq \frac{2L_{h^*}}{t}\operatorname{cothc}\left(\frac{\sqrt{\mu}}{2}t\right)\int_0^t\|\Delta_W(s)\|\,ds. \\
&\leq \frac{2L_{h^*}}{t}\operatorname{cothc}\left(\frac{\sqrt{\mu}}{2}t\right)\int_0^t\frac{s^2}{2}A(t)\,ds \\
&= \frac{L_{h^*}2}{t}\operatorname{cothc}\left(\frac{\sqrt{\mu}}{2}t\right)A(t)\frac{t^3}{6} \\
&= \frac{L_{h^*}t^2}{3}\operatorname{cothc}\left(\frac{\sqrt{\mu}}{2}t\right)A(t).
\end{aligned}$$

Taking the supremum yields

$$B(t) \leq \frac{L_{h^*} t^2}{3} \operatorname{cothc}\left(\frac{\sqrt{\mu}}{2} t\right) A(t). \tag{81}$$

Now, combining the inequalities (80) and (81), we have

$$\begin{aligned} A(t) &\leq \frac{1}{2} \operatorname{tanhc}\left(\frac{\sqrt{\mu}}{2} t\right) \left((\mu L_h + L_f) B(t) + \frac{\mu t^2}{2} A(t) \right) \\ &\leq \frac{1}{2} \operatorname{tanhc}\left(\frac{\sqrt{\mu}}{2} t\right) \left((\mu L_h + L_f) \frac{L_{h^*} t^2}{3} \operatorname{cothc}\left(\frac{\sqrt{\mu}}{2} t\right) + \frac{\mu t^2}{2} \right) A(t). \end{aligned}$$

Using continuity, it is easy to see that there exists a constant $T_{\text{small}} > 0$ such that the following inequality holds whenever $t \in (0, T_{\text{small}})$:

$$\frac{1}{2} \operatorname{tanhc}\left(\frac{\sqrt{\mu}}{2} t\right) \left((\mu L_h + L_f) \frac{L_{h^*} t^2}{3} \operatorname{cothc}\left(\frac{\sqrt{\mu}}{2} t\right) + \frac{\mu t^2}{2} \right) < 1.$$

Thus, for $t \in (0, T_{\text{small}})$, we have $A(t) \leq 1 \cdot A(t)$, which implies $A(t) = 0$ because $A(t)$ is nonnegative by its definition. Finally, $B(t) = 0$ follows from (81). This completes the proof.

D.4. Proof of Theorem 4.1

Substituting $h(x) = \frac{1}{2}\|x\|^2$, $\alpha(t) = \log(\frac{2}{t} \operatorname{cothc}(\frac{\sqrt{\mu}}{2} t))$, and $\beta(t) = \log(\frac{t^2}{4} \operatorname{sinhc}^2(\frac{\sqrt{\mu}}{2} t))$ in the Lyapunov function (63) gives

$$V(X, Z, t) = \frac{1}{2} \cosh^2\left(\frac{\sqrt{\mu}}{2} t\right) \|Z - x^*\|^2 + \frac{t^2}{4} \operatorname{sinhc}^2\left(\frac{\sqrt{\mu}}{2} t\right) \left(f(X) - f\left(x^*\right)\right). \tag{82}$$

Also, the energy function (64) can be written as

$$\mathcal{E}(t) = V(X(t), Z(t), t) = \frac{1}{2} \cosh^2\left(\frac{\sqrt{\mu}}{2} t\right) \|Z(t) - x^*\|^2 + \frac{t^2}{4} \operatorname{sinhc}^2\left(\frac{\sqrt{\mu}}{2} t\right) \left(f(X(t)) - f\left(x^*\right)\right). \tag{83}$$

Beccuase $\mathcal{E}(t)$ is monotonically non-increasing (see Appendix C.3), we have $\mathcal{E}(t) \leq \mathcal{E}(0)$. Writing this inequality explicitly, we obtain

$$f(X(t)) - f(x^*) \leq \frac{2}{t^2} \operatorname{cschc}^2\left(\frac{\sqrt{\mu}}{2} t\right) \|x_0 - x^*\|^2.$$

Since cschc^2 is decreasing on $[0, \infty)$, this implies that the unified AGM ODE (8) achieves an $O(1/t^2)$ convergence rate regardless of the value of $\mu \geq 0$. When $\mu > 0$, since $\frac{1}{t^2} \operatorname{cschc}^2\left(\frac{\sqrt{\mu}}{2} t\right) \sim \mu e^{-\sqrt{\mu} t}$ as $t \to \infty$, the unified AGM ODE achieves an $O(e^{-\sqrt{\mu} t})$ convergence rate. Combining these bounds, we conclude that the unified AGM ODE achieves an

$$O\left(\min\left\{1/t^2, e^{-\sqrt{\mu} t}\right\}\right)$$

convergence rate.

D.5. Proof of Theorem 4.2

Note that the unified AGM is equivalent to the unified AGM family (68) with $\mathbf{t}_k := \iota \sqrt{s} k$. For this sequence $(\mathbf{t}_k)$, we can check that the following conditions hold (see Appendix D.5.1):

$$\frac{2\sqrt{s}}{\mathbf{t}_k} \operatorname{cothc}\left(\frac{\sqrt{\mu}}{2} \mathbf{t}_k\right) \leq 1 \text{ for } k \geq 2 \tag{84}$$

and

$$\left(1 - \frac{2\sqrt{s}}{\mathbf{t}_{k+1}} \operatorname{cothc}\left(\frac{\sqrt{\mu}}{2} \mathbf{t}_{k+1}\right)\right) \frac{\mathbf{t}_{k+1}^2}{4} \operatorname{sinhc}^2\left(\frac{\sqrt{\mu}}{2} \mathbf{t}_{k+1}\right) \leq \frac{\mathbf{t}_k^2}{4} \operatorname{sinhc}^2\left(\frac{\sqrt{\mu}}{2} \mathbf{t}_k\right) \text{ for } k \geq 0. \tag{85}$$

Now, we claim that the following discrete-time energy function is non-increasing:

$$\mathcal{E}_k = V\left(x_k, z_k, \mathbf{t}_k\right) = \frac{1}{2} \cosh^2\left(\frac{\sqrt{\mu}}{2} \mathbf{t}_k\right) \|z_k - x^*\|^2 + \frac{\mathbf{t}_k^2}{4} \operatorname{sinhc}^2\left(\frac{\sqrt{\mu}}{2} \mathbf{t}_k\right) \left(f\left(x_k\right) - f\left(x^*\right)\right), \tag{86}$$

where the Lyapunov function V is defined in (82).

Note that when $\mu > 0$, the inequality (85) can be written as

$$\begin{aligned}
0 &\geq \left(1-\sqrt{\mu s}\coth\left(\frac{\sqrt{\mu}}{2}\mathbf{t}_{k+1}\right)\right)\frac{1}{\mu}\sinh^2\left(\frac{\sqrt{\mu}}{2}\mathbf{t}_{k+1}\right)-\frac{1}{\mu}\sinh^2\left(\frac{\sqrt{\mu}}{2}\mathbf{t}_{k}\right)\\
&=\frac{1}{\mu}\sinh^2\left(\frac{\sqrt{\mu}}{2}\mathbf{t}_{k+1}\right)-\sqrt{\frac{s}{\mu}}\sinh\left(\frac{\sqrt{\mu}}{2}\mathbf{t}_{k+1}\right)\cosh\left(\frac{\sqrt{\mu}}{2}\mathbf{t}_{k+1}\right)-\frac{1}{\mu}\sinh^2\left(\frac{\sqrt{\mu}}{2}\mathbf{t}_{k}\right)\\
&=\frac{1}{\mu}\cosh^2\left(\frac{\sqrt{\mu}}{2}\mathbf{t}_{k+1}\right)-\sqrt{\frac{s}{\mu}}\sinh\left(\frac{\sqrt{\mu}}{2}\mathbf{t}_{k+1}\right)\cosh\left(\frac{\sqrt{\mu}}{2}\mathbf{t}_{k+1}\right)-\frac{1}{\mu}\cosh^2\left(\frac{\sqrt{\mu}}{2}\mathbf{t}_{k}\right)\\
&=\left(1-\sqrt{\mu s}\tanh\left(\frac{\sqrt{\mu}}{2}\mathbf{t}_{k+1}\right)\right)\frac{1}{\mu}\cosh^2\left(\frac{\sqrt{\mu}}{2}\mathbf{t}_{k+1}\right)-\frac{1}{\mu}\cosh^2\left(\frac{\sqrt{\mu}}{2}\mathbf{t}_{k}\right).
\end{aligned}$$

Thus, the following inequality holds for all $\mu \geq 0$ (it clearly holds for $\mu = 0$):

$$\left(1-\frac{\mu\sqrt{s}\mathbf{t}_{k+1}}{2}\operatorname{tanhc}\left(\frac{\sqrt{\mu}}{2}\mathbf{t}_{k+1}\right)\right)\cosh^2\left(\frac{\sqrt{\mu}}{2}\mathbf{t}_{k+1}\right)\leq\cosh^2\left(\frac{\sqrt{\mu}}{2}\mathbf{t}_{k}\right). \tag{87}$$

Using (85) and (87), we have

$$\begin{aligned}
&\mathcal{E}_{k+1}-\mathcal{E}_k\\
&=\frac{1}{2}\cosh^2\left(\frac{\sqrt{\mu}}{2}\mathbf{t}_{k+1}\right)\|z_{k+1}-x^*\|^2-\frac{1}{2}\cosh^2\left(\frac{\sqrt{\mu}}{2}\mathbf{t}_{k}\right)\|z_k-x^*\|^2\\
&\quad+\frac{\mathbf{t}_{k+1}^2}{4}\operatorname{sinhc}^2\left(\frac{\sqrt{\mu}}{2}\mathbf{t}_{k+1}\right)(f(x_{k+1})-f(x^*))-\frac{\mathbf{t}_{k}^2}{4}\operatorname{sinhc}^2\left(\frac{\sqrt{\mu}}{2}\mathbf{t}_{k}\right)(f(x_k)-f(x^*))\\
&\leq\frac{1}{2}\cosh^2\left(\frac{\sqrt{\mu}}{2}\mathbf{t}_{k+1}\right)\|z_{k+1}-x^*\|^2-\frac{1}{2}\left(1-\frac{\mu\sqrt{s}\mathbf{t}_{k+1}}{2}\operatorname{tanhc}\left(\frac{\sqrt{\mu}}{2}\mathbf{t}_{k+1}\right)\right)\cosh^2\left(\frac{\sqrt{\mu}}{2}\mathbf{t}_{k+1}\right)\|z_k-x^*\|^2\\
&\quad+\frac{\mathbf{t}_{k+1}^2}{4}\operatorname{sinhc}^2\left(\frac{\sqrt{\mu}}{2}\mathbf{t}_{k+1}\right)(f(x_{k+1})-f(x^*))\\
&\quad-\left(1-\frac{2\sqrt{s}}{\mathbf{t}_{k+1}}\operatorname{cothc}\left(\frac{\sqrt{\mu}}{2}\mathbf{t}_{k+1}\right)\right)\frac{\mathbf{t}_{k+1}^2}{4}\operatorname{sinhc}^2\left(\frac{\sqrt{\mu}}{2}\mathbf{t}_{k+1}\right)(f(x_k)-f(x^*)).
\end{aligned}$$

Substituting

$$z_{k+1}=y_k+\left(1-\frac{\mu\sqrt{s}\mathbf{t}_{k+1}}{2}\operatorname{tanhc}\left(\frac{\sqrt{\mu}}{2}\mathbf{t}_{k+1}\right)\right)(z_k-y_k)-\frac{\sqrt{s}\mathbf{t}_{k+1}}{2}\operatorname{tanhc}\left(\frac{\sqrt{\mu}}{2}\mathbf{t}_{k+1}\right)\nabla f(y_k)$$

into the inequality above, we have

$$\begin{aligned}
&\mathcal{E}_{k+1}-\mathcal{E}_k\\
&\leq\frac{1}{2}\cosh^2\left(\frac{\sqrt{\mu}}{2}\mathbf{t}_{k+1}\right)\\
&\quad\times\left\|\left(1-\frac{\mu\sqrt{s}\mathbf{t}_{k+1}}{2}\operatorname{tanhc}\left(\frac{\sqrt{\mu}}{2}\mathbf{t}_{k+1}\right)\right)(z_k-y_k)-\frac{\sqrt{s}\mathbf{t}_{k+1}}{2}\operatorname{tanhc}\left(\frac{\sqrt{\mu}}{2}\mathbf{t}_{k+1}\right)\nabla f(y_k)-(x^*-y_k)\right\|^2\\
&\quad-\frac{1}{2}\left(1-\frac{\mu\sqrt{s}\mathbf{t}_{k+1}}{2}\operatorname{tanhc}\left(\frac{\sqrt{\mu}}{2}\mathbf{t}_{k+1}\right)\right)\cosh^2\left(\frac{\sqrt{\mu}}{2}\mathbf{t}_{k+1}\right)\|(z_k-y_k)-(x^*-y_k)\|^2\\
&\quad+\frac{\mathbf{t}_{k+1}^2}{4}\operatorname{sinhc}^2\left(\frac{\sqrt{\mu}}{2}\mathbf{t}_{k+1}\right)(f(x_{k+1})-f(x^*))\\
&\quad-\left(1-\frac{2\sqrt{s}}{\mathbf{t}_{k+1}}\operatorname{cothc}\left(\frac{\sqrt{\mu}}{2}\mathbf{t}_{k+1}\right)\right)\frac{\mathbf{t}_{k+1}^2}{4}\operatorname{sinhc}^2\left(\frac{\sqrt{\mu}}{2}\mathbf{t}_{k+1}\right)(f(x_k)-f(x^*))\\
&=\frac{1}{2}\cosh^2\left(\frac{\sqrt{\mu}}{2}\mathbf{t}_{k+1}\right)\left(\left(1-\frac{\mu\sqrt{s}\mathbf{t}_{k+1}}{2}\operatorname{tanhc}\left(\frac{\sqrt{\mu}}{2}\mathbf{t}_{k+1}\right)\right)^2-\left(1-\frac{\mu\sqrt{s}\mathbf{t}_{k+1}}{2}\operatorname{tanhc}\left(\frac{\sqrt{\mu}}{2}\mathbf{t}_{k+1}\right)\right)\right)\|z_k-y_k\|^2
\end{aligned}$$

$$
\begin{aligned}
&+ \frac{\sqrt{s}\mathbf{t}_{k+1}}{2}\operatorname{sinhc}\left(\frac{\sqrt{\mu}}{2}\mathbf{t}_{k+1}\right)\cosh\left(\frac{\sqrt{\mu}}{2}\mathbf{t}_{k+1}\right)\langle\nabla f\left(y_k\right), x^* - y_k\rangle\\
&+ \frac{\mu\sqrt{s}\mathbf{t}_{k+1}}{4}\operatorname{sinhc}\left(\frac{\sqrt{\mu}}{2}\mathbf{t}_{k+1}\right)\cosh\left(\frac{\sqrt{\mu}}{2}\mathbf{t}_{k+1}\right)\left\|x^* - y_k\right\|^2\\
&- \frac{\sqrt{s}\mathbf{t}_{k+1}}{2}\operatorname{sinhc}\left(\frac{\sqrt{\mu}}{2}\mathbf{t}_{k+1}\right)\cosh\left(\frac{\sqrt{\mu}}{2}\mathbf{t}_{k+1}\right)\left(1 - \frac{\mu\sqrt{s}\mathbf{t}_{k+1}}{2}\operatorname{tanhc}\left(\frac{\sqrt{\mu}}{2}\mathbf{t}_{k+1}\right)\right)\langle\nabla f(y_k), z_k - y_k\rangle\\
&+ \frac{s\mathbf{t}_{k+1}^2}{8}\operatorname{sinhc}^2\left(\frac{\sqrt{\mu}}{2}\mathbf{t}_{k+1}\right)\left\|\nabla f\left(y_k\right)\right\|^2 + \frac{\mathbf{t}_{k+1}^2}{4}\operatorname{sinhc}^2\left(\frac{\sqrt{\mu}}{2}\mathbf{t}_{k+1}\right)\left(f\left(x_{k+1}\right) - f\left(x^*\right)\right)\\
&- \left(1 - \frac{2\sqrt{s}}{\mathbf{t}_{k+1}}\operatorname{cothc}\left(\frac{\sqrt{\mu}}{2}\mathbf{t}_{k+1}\right)\right)\frac{\mathbf{t}_{k+1}^2}{4}\operatorname{sinhc}^2\left(\frac{\sqrt{\mu}}{2}\mathbf{t}_{k+1}\right)\left(f\left(x_k\right) - f\left(x^*\right)\right).
\end{aligned}
$$

Since

$$
0 \le 1 - \sqrt{\mu s} \le 1 - \frac{\mu\sqrt{s}\mathbf{t}_{k+1}}{2}\operatorname{tanhc}\left(\frac{\sqrt{\mu}}{2}\mathbf{t}_{k+1}\right) \le 1,
$$

we have

$$
\begin{aligned}
\frac{1}{2}\cosh^2\left(\frac{\sqrt{\mu}}{2}\mathbf{t}_{k+1}\right)\Bigg(&\left(1 - \frac{\mu\sqrt{s}\mathbf{t}_{k+1}}{2}\operatorname{tanhc}\left(\frac{\sqrt{\mu}}{2}\mathbf{t}_{k+1}\right)\right)^2\\
&- \left(1 - \frac{\mu\sqrt{s}\mathbf{t}_{k+1}}{2}\operatorname{tanhc}\left(\frac{\sqrt{\mu}}{2}\mathbf{t}_{k+1}\right)\right)\Bigg)\left\|z_k - y_k\right\|^2 \le 0.
\end{aligned}
$$

Therefore, we deduce that

$$
\begin{aligned}
&\mathcal{E}_{k+1} - \mathcal{E}_k\\
&\le \frac{\sqrt{s}\mathbf{t}_{k+1}}{2}\operatorname{sinhc}\left(\frac{\sqrt{\mu}}{2}\mathbf{t}_{k+1}\right)\cosh\left(\frac{\sqrt{\mu}}{2}\mathbf{t}_{k+1}\right)\langle\nabla f\left(y_k\right), x^* - y_k\rangle\\
&\quad + \frac{\mu\sqrt{s}\mathbf{t}_{k+1}}{4}\operatorname{sinhc}\left(\frac{\sqrt{\mu}}{2}\mathbf{t}_{k+1}\right)\cosh\left(\frac{\sqrt{\mu}}{2}\mathbf{t}_{k+1}\right)\left\|x^* - y_k\right\|^2\\
&\quad - \frac{\sqrt{s}\mathbf{t}_{k+1}}{2}\operatorname{sinhc}\left(\frac{\sqrt{\mu}}{2}\mathbf{t}_{k+1}\right)\cosh\left(\frac{\sqrt{\mu}}{2}\mathbf{t}_{k+1}\right)\\
&\qquad \times\left(1 - \frac{\mu\sqrt{s}\mathbf{t}_{k+1}}{2}\operatorname{tanhc}\left(\frac{\sqrt{\mu}}{2}\mathbf{t}_{k+1}\right)\right)\langle\nabla f(y_k), z_k - y_k\rangle\\
&\quad + \frac{s\mathbf{t}_{k+1}^2}{8}\operatorname{sinhc}^2\left(\frac{\sqrt{\mu}}{2}\mathbf{t}_{k+1}\right)\left\|\nabla f\left(y_k\right)\right\|^2\\
&\quad + \frac{\mathbf{t}_{k+1}^2}{4}\operatorname{sinhc}^2\left(\frac{\sqrt{\mu}}{2}\mathbf{t}_{k+1}\right)\left(f\left(x_{k+1}\right) - f\left(x^*\right)\right)\\
&\quad - \left(1 - \frac{2\sqrt{s}}{\mathbf{t}_{k+1}}\operatorname{cothc}\left(\frac{\sqrt{\mu}}{2}\mathbf{t}_{k+1}\right)\right)\frac{\mathbf{t}_{k+1}^2}{4}\operatorname{sinhc}^2\left(\frac{\sqrt{\mu}}{2}\mathbf{t}_{k+1}\right)\left(f\left(x_k\right) - f\left(x^*\right)\right).
\end{aligned}
$$

Now, in order to show that $\mathcal{E}_k$ is non-increasing, it suffices to show that the right-hand side (RHS) of the inequality above is non-positive. By the μ-strong convexity of f, we have

$$
0 \ge f\left(y_k\right) - f\left(x^*\right) + \langle\nabla f\left(y_k\right), x^* - y_k\rangle + \frac{\mu}{2}\left\|x^* - y_k\right\|^2.
$$

Moreover, it follows from the convexity and the $\frac{1}{s}$-smoothness of f that

$$
0 \ge f\left(y_k\right) - f\left(x_k\right) + \langle\nabla f(y_k), x_k - y_k\rangle
$$

and

$$
0 \ge f(x_{k+1}) - f(y_k) + \frac{s}{2}\left\|\nabla f(y_k)\right\|^2,
$$

respectively. Note that

$$x_k - y_k = -\frac{\tau_k}{1-\tau_k}\left(z_k - y_k\right) = -\frac{\frac{2\sqrt{s}}{\mathbf{t}_{k+1}}\operatorname{cothc}\left(\frac{\sqrt{\mu}}{2}\mathbf{t}_{k+1}\right) - \mu s}{1 - \frac{2\sqrt{s}}{\mathbf{t}_{k+1}}\operatorname{cothc}\left(\frac{\sqrt{\mu}}{2}\mathbf{t}_{k+1}\right)}\left(z_k - y_k\right).$$

Taking a weighted sum of the inequalities above yields (the condition (84) ensures that these weights are non-negative for $k \geq 1$, and the case $k = 0$ is trivial because $y_0 = x_0$)

$$\begin{aligned}
0 \geq\ & \frac{2\sqrt{s}}{\mathbf{t}_{k+1}}\operatorname{cothc}\left(\frac{\sqrt{\mu}}{2}\mathbf{t}_{k+1}\right)\frac{\mathbf{t}_{k+1}^2}{4}\operatorname{sinhc}^2\left(\frac{\sqrt{\mu}}{2}\mathbf{t}_{k+1}\right) \\
& \quad \times \left[f\left(y_k\right) - f\left(x^*\right) + \left\langle \nabla f\left(y_k\right), x^* - y_k\right\rangle + \frac{\mu}{2}\left\|x^* - y_k\right\|^2\right] \\
& + \left(1 - \frac{2\sqrt{s}}{\mathbf{t}_{k+1}}\operatorname{cothc}\left(\frac{\sqrt{\mu}}{2}\mathbf{t}_{k+1}\right)\right)\frac{\mathbf{t}_{k+1}^2}{4}\operatorname{sinhc}^2\left(\frac{\sqrt{\mu}}{2}\mathbf{t}_{k+1}\right) \\
& \quad \times \left[f\left(y_k\right) - f\left(x_k\right) + \langle \nabla f(y_k), x_k - y_k\rangle\right] \\
& + \frac{\mathbf{t}_{k+1}^2}{4}\operatorname{sinhc}^2\left(\frac{\sqrt{\mu}}{2}\mathbf{t}_{k+1}\right)\left[f(x_{k+1}) - f(y_k) + \frac{s}{2}\left\|\nabla f(y_k)\right\|^2\right] \\
=\ & \frac{\sqrt{s}\mathbf{t}_{k+1}}{2}\operatorname{sinhc}\left(\frac{\sqrt{\mu}}{2}\mathbf{t}_{k+1}\right)\cosh\left(\frac{\sqrt{\mu}}{2}\mathbf{t}_{k+1}\right)\left\langle \nabla f\left(y_k\right), x^* - y_k\right\rangle \\
& + \frac{\mu\sqrt{s}\mathbf{t}_{k+1}}{4}\operatorname{sinhc}\left(\frac{\sqrt{\mu}}{2}\mathbf{t}_{k+1}\right)\cosh\left(\frac{\sqrt{\mu}}{2}\mathbf{t}_{k+1}\right)\left\|x^* - y_k\right\|^2 \\
& - \left(\frac{2\sqrt{s}}{\mathbf{t}_{k+1}}\operatorname{cothc}\left(\frac{\sqrt{\mu}}{2}\mathbf{t}_{k+1}\right) - \mu s\right)\frac{\mathbf{t}_{k+1}^2}{4}\operatorname{sinhc}^2\left(\frac{\sqrt{\mu}}{2}\mathbf{t}_{k+1}\right)\left\langle \nabla f(y_k), z_k - y_k\right\rangle \\
& + \frac{s\mathbf{t}_{k+1}^2}{8}\operatorname{sinhc}^2\left(\frac{\sqrt{\mu}}{2}\mathbf{t}_{k+1}\right)\left\|\nabla f\left(y_k\right)\right\|^2 \\
& + \frac{2\sqrt{s}}{\mathbf{t}_{k+1}}\operatorname{cothc}\left(\frac{\sqrt{\mu}}{2}\mathbf{t}_{k+1}\right)\frac{\mathbf{t}_{k+1}^2}{4}\operatorname{sinhc}^2\left(\frac{\sqrt{\mu}}{2}\mathbf{t}_{k+1}\right)\left(f\left(y_k\right) - f\left(x^*\right)\right) \\
& + \left(1 - \frac{2\sqrt{s}}{\mathbf{t}_{k+1}}\operatorname{cothc}\left(\frac{\sqrt{\mu}}{2}\mathbf{t}_{k+1}\right)\right)\frac{\mathbf{t}_{k+1}^2}{4}\operatorname{sinhc}^2\left(\frac{\sqrt{\mu}}{2}\mathbf{t}_{k+1}\right)\left(f\left(y_k\right) - f\left(x_k\right)\right) \\
& + \frac{\mathbf{t}_{k+1}^2}{4}\operatorname{sinhc}^2\left(\frac{\sqrt{\mu}}{2}\mathbf{t}_{k+1}\right)\left(f(x_{k+1}) - f(y_k)\right) \\
=\ & \frac{\sqrt{s}\mathbf{t}_{k+1}}{2}\operatorname{sinhc}\left(\frac{\sqrt{\mu}}{2}\mathbf{t}_{k+1}\right)\cosh\left(\frac{\sqrt{\mu}}{2}\mathbf{t}_{k+1}\right)\left\langle \nabla f\left(y_k\right), x^* - y_k\right\rangle \\
& + \frac{\mu\sqrt{s}\mathbf{t}_{k+1}}{4}\operatorname{sinhc}\left(\frac{\sqrt{\mu}}{2}\mathbf{t}_{k+1}\right)\cosh\left(\frac{\sqrt{\mu}}{2}\mathbf{t}_{k+1}\right)\left\|x^* - y_k\right\|^2 \\
& - \frac{\sqrt{s}\mathbf{t}_{k+1}}{2}\operatorname{sinhc}\left(\frac{\sqrt{\mu}}{2}\mathbf{t}_{k+1}\right)\cosh\left(\frac{\sqrt{\mu}}{2}\mathbf{t}_{k+1}\right) \\
& \quad \times \left(1 - \frac{\mu\sqrt{s}\mathbf{t}_{k+1}}{2}\operatorname{tanhc}\left(\frac{\sqrt{\mu}}{2}\mathbf{t}_{k+1}\right)\right)\left\langle \nabla f(y_k), z_k - y_k\right\rangle \\
& + \frac{s\mathbf{t}_{k+1}^2}{8}\operatorname{sinhc}^2\left(\frac{\sqrt{\mu}}{2}\mathbf{t}_{k+1}\right)\left\|\nabla f\left(y_k\right)\right\|^2 \\
& + \frac{\mathbf{t}_{k+1}^2}{4}\operatorname{sinhc}^2\left(\frac{\sqrt{\mu}}{2}\mathbf{t}_{k+1}\right)\left(f\left(x_{k+1}\right) - f\left(x^*\right)\right) \\
& - \left(1 - \frac{2\sqrt{s}}{\mathbf{t}_{k+1}}\operatorname{cothc}\left(\frac{\sqrt{\mu}}{2}\mathbf{t}_{k+1}\right)\right)\frac{\mathbf{t}_{k+1}^2}{4}\operatorname{sinhc}^2\left(\frac{\sqrt{\mu}}{2}\mathbf{t}_{k+1}\right)\left(f\left(x_k\right) - f\left(x^*\right)\right).
\end{aligned}$$

Thus, the energy function (86) is non-increasing. Writing $\mathcal{E}_k \leq \mathcal{E}_0$ explicitly, we obtain

$$\begin{aligned} f(x_k) - f(x^*) &\leq \frac{4}{\mathbf{t}_k^2} \operatorname{cschc}^2\left(\frac{\sqrt{\mu}}{2}\mathbf{t}_k\right)\left(\frac{1}{2}\cosh^2\left(\frac{\sqrt{\mu}}{2}\mathbf{t}_0\right)\|x_0 - x^*\|^2 + \frac{\mathbf{t}_0^2}{4}\operatorname{sinhc}^2\left(\frac{\sqrt{\mu}}{2}\mathbf{t}_0\right)(f(x_0) - f(x^*))\right) \\ &= \frac{2}{\iota^2 s k^2}\operatorname{cschc}^2\left(\frac{\iota\sqrt{\mu s}}{2}k\right)\|x_0 - x^*\|^2. \end{aligned}$$

Because cschc^2 is decreasing on $[0, \infty)$, we have

$$\frac{2}{\iota^2 s k^2}\operatorname{cschc}^2\left(\frac{\iota\sqrt{\mu s}}{2}k\right) \leq \frac{2}{\iota^2 s k^2}.$$

This implies that the convergence guarantee of the unified AGM is always better than that of AGM-C and that the unified AGM achieves an $O(1/k^2)$ convergence rate, regardless of the value of μ. When $\mu > 0$, since

$$\frac{4}{\mathbf{t}_k^2}\operatorname{cschc}^2\left(\frac{\sqrt{\mu}}{2}\mathbf{t}_k\right) \sim 4\mu e^{-\sqrt{\mu}\mathbf{t}_k} = 4\mu\left(1 - \sqrt{\mu s}\right)^k \text{ as } k \to \infty,$$

the unified AGM achieves an $O((1 - \sqrt{\mu s})^k)$ convergence rate. Combining these two guarantees, we conclude that the unified AGM achieves an

$$O\left(\min\left\{1/k^2, \left(1 - \sqrt{\mu s}\right)^k\right\}\right)$$

convergence rate. This completes the proof.

D.5.1. Checking Conditions on Time Sequence

We show that the sequence $\mathbf{t}_k := \iota\sqrt{s}k$ satisfies the conditions (84) and (85). For convenience, we assume $\mu > 0$ (the case $\mu = 0$ can be handled easily). The condition (84) follows from

$$\begin{aligned} \frac{2\sqrt{s}}{\mathbf{t}_k}\operatorname{cothc}\left(\frac{\sqrt{\mu}}{2}\mathbf{t}_k\right) &= \sqrt{\mu s}\coth\left(\frac{\sqrt{\mu}}{2}\mathbf{t}_k\right) \\ &\leq \sqrt{\mu s}\coth\left(\frac{\sqrt{\mu}}{2}\mathbf{t}_2\right) \\ &= \sqrt{\mu s}\coth\left(-\log\left(1 - \sqrt{\mu s}\right)\right) \\ &= \sqrt{\mu s}\frac{1 + e^{2\log\left(1-\sqrt{\mu s}\right)}}{1 - e^{2\log\left(1-\sqrt{\mu s}\right)}} \\ &= \sqrt{\mu s}\frac{1 + \left(1 - \sqrt{\mu s}\right)^2}{1 - \left(1 - \sqrt{\mu s}\right)^2} \\ &\leq 1, \end{aligned}$$

where the last inequality holds because $\sqrt{\mu s} \in (0, 1)$. To prove (85), it suffices to show that the inequality

$$\sinh^2\left(\frac{\sqrt{\mu}}{2}t\right) - \sqrt{\mu s}\sinh\left(\frac{\sqrt{\mu}}{2}t\right)\cosh\left(\frac{\sqrt{\mu}}{2}t\right) - \sinh^2\left(\frac{\sqrt{\mu}}{2}t + \frac{1}{2}\log\left(1 - \sqrt{\mu s}\right)\right) \leq 0$$

holds for all $t \in \mathbb{R}$. Letting $r = e^{\frac{\sqrt{\mu}}{2}t}$, this inequality can be expressed as

$$\frac{r^2 + r^{-2} - 2}{4} - \sqrt{\mu s}\frac{r^2 - r^{-2}}{4} - \frac{\left(1 - \sqrt{\mu s}\right)r^2 + \left(1 - \sqrt{\mu s}\right)^{-1}r^{-2} - 2}{4} \leq 0.$$

Letting $q = r^2$ and multiplying both sides by $4q$, the inequality can be rewritten as

$$\begin{aligned} 0 &\geq q^2 + 1 - 2q - \sqrt{\mu s}\left(q^2 - 1\right) - \left(1 - \sqrt{\mu s}\right)q^2 - \left(1 - \sqrt{\mu s}\right)^{-1} + 2q \\ &= 1 + \sqrt{\mu s} - \frac{1}{1 - \sqrt{\mu s}} \\ &= \frac{-\mu s}{1 - \sqrt{\mu s}}, \end{aligned}$$

which clearly holds.

D.6. Unified AGM Converges to Unified AGM ODE

Note that the unified AGM family (68) is the three-sequence scheme (41) with

$$\begin{aligned}\tau_k &= \frac{\frac{2\sqrt{s}}{\mathbf{t}_{k+1}}\operatorname{cothc}\left(\frac{\sqrt{\mu}}{2}\mathbf{t}_{k+1}\right) - \mu s}{1-\mu s}\\ \delta_k &= \frac{\sqrt{s}\mathbf{t}_{k+1}}{2}\operatorname{tanhc}\left(\frac{\sqrt{\mu}}{2}\mathbf{t}_{k+1}\right).\end{aligned} \tag{88}$$

If the sequence $(\mathbf{t}_k)_{k=0}^{\infty}$ in $[0,\infty)$ satisfies the conditions (42) and (43), then we have

$$\begin{aligned}\lim_{s\to 0}\frac{\tau_{\mathbf{k}(t)}}{\sqrt{s}} &= \frac{2}{t}\operatorname{cothc}\left(\frac{\sqrt{\mu}}{2}t\right)\\ \lim_{s\to 0}\frac{\delta_{\mathbf{k}(t)}}{\sqrt{s}} &= \lim_{s\to 0}\frac{t}{2}\operatorname{tanhc}\left(\frac{\sqrt{\mu}}{2}t\right) = \frac{t}{2}\operatorname{tanhc}\left(\frac{\sqrt{\mu}}{2}t\right)\end{aligned}$$

for all $t>0$, where $\mathbf{k}$ is the inverse function of the sequence $\mathbf{t}$. In this case, the result in Appendix B.2.1 implies that the unified AGM family (68) converges to the unified AGM system (69) as $s\to 0$.

Because the sequence $\mathbf{t}_k = \iota\sqrt{s}k$ clearly satisfies the conditions (42) and (43) and the unified AGM is equivalent to the unified AGM family (68) with $\mathbf{t}_k = \iota\sqrt{s}k$, the unified AGM converges to the unified AGM system (69), which is equivalent to the unified AGM ODE.

D.7. AGM-SC is Asymptotic Limit of Unified AGM

The following proposition rigorously states that AGM-SC is the asymptotic limit of the unified AGM.

Proposition D.2. *Let $x_0\in\mathbb{R}^n$ and $K\in\mathbb{N}$. Then, for every real number $\epsilon>0$, there exists a positive integer R such that*

$$\|x^1_{k_0+K} - x^2_{k_0+K}\| \le \epsilon \text{ for every integer } k_0 \ge R,$$

where

- *The iterates (x^1,z^1) are generated by the unified AGM with the initial point $x^1_{k_0}=z^1_{k_0}$ at the initial iteration k_0 (that is, we run Algorithm 1 where the for loop is started at $k=k_0$ instead of $k=0$).*
- *The iterates (x^2,z^2) are generated by AGM-SC with the initial point $x^2_{k_0}=z^2_{k_0}$ at the initial iteration k_0.*

Proof. Let $\tau_k = \frac{1}{1-q}\left(\frac{2}{\iota(k+1)}\operatorname{cothc}\left(\frac{k+1}{2}\iota\sqrt{q}\right) - q\right)$, $\tau = \frac{\sqrt{q}}{1+\sqrt{q}}$, $\delta_k = \frac{\iota s(k+1)}{2}\operatorname{tanhc}\left(\frac{k+1}{2}\iota\sqrt{q}\right)$, and $\delta = \sqrt{\frac{s}{\mu}}$. Then, it is easy to check that $\tau_k\to\tau$ and $\delta_k\to\delta$ as $k\to\infty$. In addition, we can write the unified AGM and AGM-SC as

$$\begin{aligned}y_k &= x_k + \tau_k(z_k - x_k)\\ x_{k+1} &= y_k - s\nabla f(y_k)\\ z_{k+1} &= z_k + \delta_k(\mu y_k - \mu z_k - \nabla f(y_k))\end{aligned}$$

and

$$\begin{aligned}y_k &= x_k + \tau(z_k - x_k)\\ x_{k+1} &= y_k - s\nabla f(y_k)\\ z_{k+1} &= z_k + \delta(\mu y_k - \mu z_k - \nabla f(y_k)),\end{aligned}$$

respectively. Now, a straightforward calculation yields

$$\begin{aligned}y^1_{k+1} - y^2_{k+1} = (1-\tau)\left(y^1_k - y^2_k - s\nabla f(y^1_k) + s\nabla f(y^2_k)\right) + \tau\left(z^1_{k+1} - z^2_{k+1}\right)\\ -(\tau_k-\tau)\,y^1_k + s\,(\tau_k-\tau)\,\nabla f(y^1_k) + (\tau_k-\tau)\,z^1_{k+1}\end{aligned}$$

and

$$z_{k+1}^1 - z_{k+1}^2 = (1-\mu\delta)\left(z_k^1 - z_k^2\right) + \mu\delta\left(y_k^1 - y_k^2\right) - \delta\left(\nabla f(y_k^1) - \nabla f(y_k^2)\right) \\ - \mu\left(\delta_k - \delta\right) z_k^1 + \mu\left(\delta_k - \delta\right) y_k^1 - \left(\delta_k - \delta\right)\nabla f(y_k^1).$$

We can show that $\|y_k\|$ and $\|z_k\|$ are bounded above by a constant which is independent of the initial iteration k_0.[20] Define $\Delta_k = (y_k^1 - y_k^2, z_k^1 - z_k^2) \in \mathbb{R}^{2n}$ and $u_k = \|\Delta_k\|$. Let $C = 1 + (s - s\tau + \delta)L_f$, where L_f is a Lipschitz continuity parameter of ∇f. Let $d = \frac{C-1}{C^K-1}\epsilon$. Then, because $\tau_k \to \tau$ and $\delta_k \to \delta$ as $k \to \infty$, we can show that

$$\text{There exists } R \in \mathbb{N} \text{ such that } k \geq k_0 \geq R \Rightarrow u_{k+1} \leq Cu_k + d.$$

Note that $u_{k_0} = 0$. Now, it is easy to show that

$$\begin{aligned} u_{k_0+K} &\leq C^K u_{k_0} + \left(1 + C + \cdots + C^{K-1}\right) d \\ &= C^K u_{k_0} + \frac{C^K - 1}{C - 1} d \\ &= \frac{C^K - 1}{C - 1} d \\ &= \epsilon \end{aligned}$$

for all $k_0 \geq R$. This completes the proof. □

D.8. Nesterov's Constant Step Scheme as Rate-Matching Discretization of Unified AGM ODE

In Appendix D.8.1, we provide a rate-matching discretization of the unified AGM system (69) with an adaptive timestep. In Appendix D.8.3, we show that this algorithm is equivalent to the *constant step scheme I* (Nesterov, 2018, Equation 2.2.19).

D.8.1. A Rate-Matching Discretization of Unified AGM ODE with Adaptive Timestep

Define the sequence $(\mathbf{t}_k)_{k=0}^\infty$ as

$$\mathbf{t}_{k+1} = \begin{cases} \text{Given constant } \mathbf{t}_0 > 0 \text{ (possibly depending on } s), & k+1 = 0 \\ \text{The largest real number satisfying (85)}, & k+1 \geq 1. \end{cases} \tag{89}$$

Then, it is easy to check that the sequence $(\mathbf{t}_k)_{k=0}^\infty$ is well-defined and strictly increasing. We refer to the unified AGM family (68) with this time sequence as the *adaptive timestep scheme*. Note that the conditions (84) and (85) hold by construction.[21] Therefore, the discrete-time energy function (86) is non-increasing for the iterates of the adaptive timestep scheme. We will show that for the sequence $(\mathbf{t}_k)$ defined by (89),

- The sequence $(\mathbf{t}_k)$ is well-defined.
- The conditions (42) and (43) hold when $\lim_{s\to 0} \mathbf{t}_0 = 0$.

Then, these results imply that if $\lim_{s\to 0} \mathbf{t}_0 = 0$, then the adaptive timestep scheme converges to the unified AGM system (69) as $s \to 0$ by the result in Appendix A.1. Because the discrete-time energy function (86) for the adaptive timestep scheme and the continuous-time energy function (83) for the unified AGM system are equivalent under the identifications $t \leftrightarrow \mathbf{t}_k$, $X(\mathbf{t}_k) \leftrightarrow x_k$, and $Z(\mathbf{t}_k) \leftrightarrow z_k$, we conclude that the adaptive timestep scheme is a rate-matching discretization of the unified AGM ODE.

[20] This can be proven by bounding $\|x_k - x^*\|$ and $\|z_k - x^*\|$ using the strong convexity of f and the fact that the energy function (86) is non-increasing after $k = k_0$. We omit the details.

[21] The first condition follows from the facts that (84) holds for the sequence $\mathbf{t}_k = k\delta$, and that we have $\mathbf{t}_k > 2\delta$ for $k \geq 2$, for the sequence $(\mathbf{t}_k)$ defined in (89).

The sequence $(\mathbf{t}_k)$ is well-defined. Because

$$\frac{4}{\mathbf{t}_{k+1}^2}\operatorname{cschc}^2\left(\frac{\sqrt{\mu}}{2}\mathbf{t}_{k+1}\right)+\mu=\frac{4}{\mathbf{t}_{k+1}^2}\operatorname{cothc}^2\left(\frac{\sqrt{\mu}}{2}\mathbf{t}_{k+1}\right),$$

the updating rule (89) is equivalent to

$$\frac{4}{\mathbf{t}_{k+1}^2}\operatorname{cothc}^2\left(\frac{\sqrt{\mu}}{2}\mathbf{t}_{k+1}\right)=\left(1-\frac{2\sqrt{s}}{\mathbf{t}_{k+1}}\operatorname{cothc}\left(\frac{\sqrt{\mu}}{2}\mathbf{t}_{k+1}\right)\right)\frac{4}{\mathbf{t}_k^2}\operatorname{cothc}^2\left(\frac{\sqrt{\mu}}{2}\mathbf{t}_k\right)$$
$$+\frac{2\mu\sqrt{s}}{\mathbf{t}_{k+1}}\operatorname{cothc}\left(\frac{\sqrt{\mu}}{2}\mathbf{t}_{k+1}\right),\ \mathbf{t}_{k+1}>0. \quad (90)$$

Introduce a sequence $(\alpha_k)_{k=-1}^{\infty}$ such that $\alpha_k=\frac{2\sqrt{s}}{\mathbf{t}_{k+1}}\operatorname{cothc}\left(\frac{\sqrt{\mu}}{2}\mathbf{t}_{k+1}\right)$. As $t\mapsto\frac{2\sqrt{s}}{t}\operatorname{cothc}\left(\frac{\sqrt{\mu}}{2}t\right)$ is a bijective map from $(0,\infty)$ to $\left(\sqrt{\mu s},\infty\right)$, the sequences $(\mathbf{t}_k)$ and (α_k) have a one-to-one relationship. Thus, the updating rule (90) is equivalent to

$$\alpha_k^2=(1-\alpha_k)\,\alpha_{k-1}^2+\mu s\alpha_k,\quad \alpha_k>\sqrt{\mu s}, \quad (91)$$

which admits a unique solution in $(\sqrt{\mu s},\infty)$ when $\alpha_{k-1}>\sqrt{\mu s}$. Thus, the sequence $(\mathbf{t}_k)$ is well-defined.

The sequence (t_k) satisfies the conditions (42) and (43). Define a function $A(t)$ as

$$A(t):=\frac{t^2}{4}\operatorname{sinhc}^2\left(\frac{\sqrt{\mu}}{2}t\right). \quad (92)$$

For $t\in(0,\infty)$, it follows from (89) that

$$\dot{A}\left(\mathbf{t}_{\mathbf{k}(t)+1}\right)=\frac{A\left(\mathbf{t}_{\mathbf{k}(t)+1}\right)-A\left(t\right)}{\sqrt{s}}=\frac{A\left(\mathbf{t}_{\mathbf{k}(t)+1}\right)-A\left(t\right)}{\mathbf{t}_{\mathbf{k}(t)+1}-t}\frac{\mathbf{t}_{\mathbf{k}(t)+1}-t}{\sqrt{s}}.$$

Because $\mathbf{t}_{\mathbf{k}(t)+1}\to t$ as $s\to 0$, taking the limit $s\to 0$ in the equation above yields

$$1=\lim_{s\to 0}\frac{\mathbf{t}_{\mathbf{k}(t)+1}-t}{\sqrt{s}}.$$

Thus, the condition (85) holds.

D.8.2. Nesterov's Constant Step Scheme

For μ-strongly (possibly with $\mu=0$) convex objective functions, Nesterov considered the following algorithm: Given an initial point $x_0=z_0\in\mathbb{R}^n$ and $\gamma_0>0$, the *constant step scheme I* (Nesterov, 2018, Equation 2.2.19) (we will also refer to this algorithm as the *original NAG*) updates the iterates as

$$\begin{aligned}\gamma_{k+1}&=(1-\alpha_k)\,\gamma_k+\mu\alpha_k\\ y_k&=\frac{1}{\gamma_k+\mu\alpha_k}\left(\alpha_k\gamma_k z_k+\gamma_{k+1}x_k\right)\\ x_{k+1}&=y_k-s\nabla f\left(y_k\right)\\ z_{k+1}&=\frac{1}{\gamma_{k+1}}\left(\left(1-\alpha_k\right)\gamma_k z_k+\mu\alpha_k y_k-\alpha_k\nabla f\left(y_k\right)\right),\end{aligned} \quad (93)$$

where the sequence $(\alpha_k)_{k=0}^{\infty}$ in $(0,1)$ is inductively defined by the equation

$$\frac{1}{s}\alpha_k^2=(1-\alpha_k)\,\gamma_k+\mu\alpha_k. \quad (94)$$

Using the estimate sequence technique, Nesterov (2018, Theorem 2.2.1) showed that the iterates of the original NAG (93) satisfy the inequality

$$f\left(x_k\right)-f\left(x^*\right)\le\left(\prod_{i=0}^{k-1}\left(1-\alpha_i\right)\right)\left(f\left(x_0\right)-f\left(x^*\right)+\frac{\gamma_0}{2}\left\|x_0-x^*\right\|^2\right) \quad (95)$$

when $s\le 1/L$.

D.8.3. Equivalence Between Adaptive Timestep Scheme and Original NAG

Surprisingly, the adaptive timestep scheme in Appendix D.8.1, which is purely obtained from the unified Lagrangian framework, is equivalent to the original NAG (93).

Proposition D.3. *The adaptive timestep scheme is equivalent to the original NAG (93) with* $\gamma_0 = \frac{4}{\mathbf{t}_0^2}\operatorname{cothc}^2\left(\frac{\sqrt{\mu}}{2}\mathbf{t}_0\right) > \mu$. *Moreover, the sequence* γ_k *and* α_k *in the original NAG can be written as* $\gamma_k = \frac{4}{\mathbf{t}_k^2}\operatorname{cothc}^2\left(\frac{\sqrt{\mu}}{2}\mathbf{t}_k\right)$ *and* $\alpha_k = \frac{2\sqrt{s}}{\mathbf{t}_{k+1}}\operatorname{cothc}\left(\frac{\sqrt{\mu}}{2}\mathbf{t}_{k+1}\right)$. *Conversely, the original NAG (93) with* $\gamma_0 > \mu$ *is equivalent to the adaptive timestep scheme, where* $\mathbf{t}_0$ *is a nonnegative constant satisfying* $\gamma_0 = \frac{4}{\mathbf{t}_0^2}\operatorname{cothc}^2\left(\frac{\sqrt{\mu}}{2}\mathbf{t}_0\right)$.

Proof. We first show that the sequences $(\alpha_k)_{k=0}^\infty$ and $(\gamma_k)_{k=0}^\infty$ generated in the original NAG (93) with $\gamma_0 = \frac{4}{\mathbf{t}_0^2}\operatorname{cothc}^2\left(\frac{\sqrt{\mu}}{2}\mathbf{t}_0\right) > \mu$ can be written as $\alpha_k = \frac{2\sqrt{s}}{\mathbf{t}_{k+1}}\operatorname{cothc}\left(\frac{\sqrt{\mu}}{2}\mathbf{t}_{k+1}\right)$ and $\gamma_k = \frac{4}{\mathbf{t}_k^2}\operatorname{cothc}^2\left(\frac{\sqrt{\mu}}{2}\mathbf{t}_k\right)$, where the sequence $(\mathbf{t}_k)_{k=0}^\infty$ is defined as (89). Note that the updating rules of (α_k) and (γ_k) in the original NAG implies

$$\gamma_{k+1} = (1-\alpha_k)\,\gamma_k + \mu\alpha_k = \frac{\alpha_k^2}{s}.$$

Thus, the updating rule for α_k (94) can be equivalently written as

$$\frac{1}{s}\alpha_k^2 = (1-\alpha_k)\,\frac{\alpha_{k-1}^2}{s} + \mu\alpha_k,$$

where we define $\alpha_{-1} := \sqrt{s\gamma_0} = \frac{2\sqrt{s}}{\mathbf{t}_0}\operatorname{cothc}\left(\frac{\sqrt{\mu}}{2}\mathbf{t}_0\right) > \sqrt{\mu s}$. This implies that the sequence $(\alpha_k)_{k=-1}^\infty$ in the original NAG and the sequence $(\alpha_k)_{k=-1}^\infty$ defined in Section D.8.1 are identical. Thus, we have $\alpha_k = \frac{2\sqrt{s}}{\mathbf{t}_{k+1}}\operatorname{cothc}\left(\frac{\sqrt{\mu}}{2}\mathbf{t}_{k+1}\right)$ and $\gamma_k = \frac{\alpha_{k-1}^2}{s} = \frac{4}{\mathbf{t}_k^2}\operatorname{cothc}^2\left(\frac{\sqrt{\mu}}{2}\mathbf{t}_k\right)$.

Now, we show that the parameters τ_k and δ_k for the original NAG are equal to those for our adaptive timestep scheme. In the original NAG, we have

$$\begin{aligned}(\alpha_k - \mu s)\,(\gamma_k + \mu\alpha_k) &= \alpha_k\gamma_k + \mu\alpha_k^2 - \mu s\gamma_k - \mu^2 s\alpha_k \\ &= \mu s\gamma_{k+1} + \alpha_k\gamma_k - \mu s\gamma_k - \mu^2 s\alpha_k \\ &= \mu s\left((1-\alpha_k)\,\gamma_k + \mu\alpha_k\right) + \alpha_k\gamma_k - \mu s\gamma_k - \mu^2 s\alpha_k \\ &= (1-\mu s)\alpha_k\gamma_k.\end{aligned}$$

Therefore, we have

$$\tau_k = \frac{\alpha_k\gamma_k}{\gamma_k + \mu\alpha_k} = \frac{\alpha_k - \mu s}{1-\mu s} = \frac{\frac{2\sqrt{s}}{\mathbf{t}_{k+1}}\operatorname{cothc}\left(\frac{\sqrt{\mu}}{2}\mathbf{t}_{k+1}\right) - \mu s}{1-\mu s}$$

and

$$\delta_k = \frac{\alpha_k}{\gamma_{k+1}} = \frac{s}{\alpha_k} = \frac{\sqrt{s}\mathbf{t}_{k+1}}{2}\operatorname{tanhc}\left(\frac{\sqrt{\mu}}{2}\mathbf{t}_{k+1}\right),$$

which are the momentum coefficients in (68). Thus, the ogirinal NAG with $\gamma_0 = \frac{4}{\mathbf{t}_0^2}\operatorname{cothc}^2\left(\frac{\sqrt{\mu}}{2}\mathbf{t}_0\right) > \mu$ is equivalent to the adaptive timestep scheme. □

The following remark shows that under the identification in Proposition D.3, the convergence rate of the adaptive timestep scheme is equivalent to the convergence rate (95) of the original NAG obtained by Nesterov (2018).

Remark D.4. Because the discrete-time energy function (86) is non-increasing, the iterates of the adaptive timestep scheme satisfy

$$\begin{aligned}
&f(x_k) - f(x^*) \\
&\leq \frac{4}{\mathbf{t}_k^2} \operatorname{cschc}^2\left(\frac{\sqrt{\mu}}{2}\mathbf{t}_k\right) \\
&\quad \times \left(\frac{1}{2}\cosh^2\left(\frac{\sqrt{\mu}}{2}\mathbf{t}_0\right)\|x_0 - x^*\|^2 + \frac{\mathbf{t}_0^2}{4}\operatorname{sinhc}^2\left(\frac{\sqrt{\mu}}{2}\mathbf{t}_0\right)(f(x_0) - f(x^*))\right) \\
&= \frac{4}{\mathbf{t}_k^2} \operatorname{cschc}^2\left(\frac{\sqrt{\mu}}{2}\mathbf{t}_k\right)\frac{\mathbf{t}_0^2}{4}\operatorname{sinhc}^2\left(\frac{\sqrt{\mu}}{2}\mathbf{t}_0\right) \\
&\quad \times \left(\frac{2}{\mathbf{t}_0^2}\operatorname{cothc}^2\left(\frac{\sqrt{\mu}}{2}\mathbf{t}_0\right)\|x_0 - x^*\|^2 + (f(x_0) - f(x^*))\right) \\
&= \prod_{i=0}^{k-1}\left(1 - \frac{2\sqrt{s}}{\mathbf{t}_{i+1}}\operatorname{cothc}\left(\frac{\sqrt{\mu}}{2}\mathbf{t}_{i+1}\right)\right)\left(\frac{2}{\mathbf{t}_0^2}\operatorname{cothc}^2\left(\frac{\sqrt{\mu}}{2}\mathbf{t}_0\right)\|x_0 - x^*\|^2 + (f(x_0) - f(x^*))\right),
\end{aligned} \tag{96}$$

where the last equality follows from our updating rule (89) of the sequence $(\mathbf{t}_k)$. Therefore, we recover the convergence rate (95) of the original NAG with $\gamma_k = \frac{4}{\mathbf{t}_k^2}\operatorname{cothc}^2\left(\frac{\sqrt{\mu}}{2}\mathbf{t}_k\right)$ and $\alpha_k = \frac{2\sqrt{s}}{\mathbf{t}_{k+1}}\operatorname{cothc}\left(\frac{\sqrt{\mu}}{2}\mathbf{t}_{k+1}\right)$.

E. Unified Higher-Order Method for Minimizing Convex and Uniformly Convex Functions

E.1. Existence and Uniqueness of Solution to Unified ATM ODE

In this subsection, we prove the existence and uniqueness of a solution to the unified ATM ODE, by using the existence and uniqueness of a solution to the system of ODEs (70) and the time-dilation property (Theorem C.1) of the unified Bregman Lagrangian flow. We first note that

- The system of ODEs (70) is the unified Bregman Lagrangian flow (5) with $\beta_1 = \log\left(\frac{t^2}{4}\operatorname{sinhc}^2\left(\frac{\sqrt{\mu}}{2}t\right)\right)$ and $\alpha_1 = \log\dot{\beta}_1$.
- The unified ATM ODE (12) is the unified Bregman Lagrangian flow (5) with $\beta_2 = p\log t + \log C + p\log\left(\operatorname{sinhc}_p\left(C^{1/p}\mu^{1/p}t\right)\right)$ and $\alpha_2 = \log\dot{\beta}_2$.

Define a function $\mathbf{T} : [0,\infty) \to [0,\infty)$ as $\mathbf{T} = \beta_1^{-1} \circ \beta_2$. Then, we have

$$\begin{aligned}
\alpha_2(t) &= \alpha_1(\mathbf{T}(t)) + \log\dot{\mathbf{T}}(t) \\
\beta_2(t) &= \beta_1(\mathbf{T}(t)).
\end{aligned}$$

Thus, by Theorem C.1, if (X_1, Z_1) is a solution to the ODE system (70), then $X_2(t) = X_1(\mathbf{T}(t))$ and $Z_2(t) = Z_1(\mathbf{T}(t))$ forms a solution to the unified ATM ODE. Thus, the existence of solution to the system (70) implies the existence of solution to the unified ATM ODE.

A similar argument shows that if (X_2, Z_2) is a solution to the unified ATM ODE, then $X_1(t) = X_2(\mathbf{T}^{-1}(t))$ and $Z_1(t) = Z_2(\mathbf{T}^{-1}(t))$ forms a solution to the system (70). It is easy to show that this correspondence is one-to-one. Thus, the uniqueness of solution to the system (70) implies the uniqueness of solution to the unified ATM ODE.

E.2. Proof of Theorem 5.1

For the unified ATM ODE (12), the Lyapunov function (63) can be written as

$$V(X,Z,t) = \cosh_p^p\left(C^{1/p}\mu^{1/p}t\right)D_h(x^*, Z) + Ct^p\operatorname{sinhc}_p^p\left(C^{1/p}\mu^{1/p}t\right)(f(X) - f(x^*)). \tag{97}$$

Thus, the proof of Theorem 3.1 (Appendix C.3) implies that the continuous-time energy function

$$\mathcal{E}(t) = V(X(t), Z(t), t) = \cosh_p^p\left(C^{1/p}\mu^{1/p}t\right)D_h(x^*, Z(t)) + Ct^p\operatorname{sinhc}_p^p\left(C^{1/p}\mu^{1/p}t\right)(f(X(t)) - f(x^*)) \tag{98}$$

is monotonically non-increasing on $[0,\infty)$. Writing $\mathcal{E}(t) \le \mathcal{E}(0)$ explicitly, we have

$$f(X(t)) - f(x^*) \le \frac{1}{Ct^p \operatorname{sinhc}_p^p\left(C^{1/p}\mu^{1/p}t\right)} D_h(x^*, x_0).$$

Since $\operatorname{sinhc}_p(0) = 1$ and sinhc_p is increasing on $[0,\infty)$ (see Appendix B.1.2), this inequality implies that the unified ATM ODE (12) achieves an $O(1/t^p)$ convergence rate regardless of the value of $\mu \ge 0$. On the other hand, when $\mu > 0$, it follows from Proposition B.1 that

$$\frac{1}{Ct^p \operatorname{sinhc}_p^p\left(C^{1/p}\mu^{1/p}t\right)} = O\left(e^{-pC^{1/p}\mu^{1/p}t}\right) \quad \text{as } t \to \infty.$$

Therefore, the unified ATM ODE achieves an $O(e^{-pC^{1/p}\mu^{1/p}t})$ convergence rate. Combining these bounds, we conclude that the unified ATM ODE achieves an

$$O\left(\min\left\{1/t^p, e^{-pC^{1/p}\mu^{1/p}t}\right\}\right)$$

convergence rate.

E.3. Proof of Theorem 5.3

We show that the discrete-time energy function

$$\mathcal{E}_k = V(x_k, z_k, \mathbf{t}_k) = (1+\mu A_k) D_h(x^*, z_k) + A_k\left(f(x_k) - f(x^*)\right) \tag{99}$$

is non-increasing, where the Lyapunov function V is defined in (97) and $\mathbf{t}_k := \sqrt[p]{A_k/C}$ if $\mu = 0$ and $\mathbf{t}_k := \operatorname{sinh}_p^{-1}(\sqrt[p]{\mu A_k})/\sqrt[p]{C\mu}$ if $\mu > 0$.

By the Bregman three-point identity[22] and the non-negativity of Bregman divergence, we have

$$\begin{aligned} D_h(x^*, z_{k+1}) &= D_h(x^*, x_{k+1}) - \langle \nabla h(z_{k+1}) - \nabla h(x_{k+1}), x^* - z_{k+1}\rangle - D_h(z_{k+1}, x_{k+1}) \\ &\le D_h(x^*, x_{k+1}) - \langle \nabla h(z_{k+1}) - \nabla h(x_{k+1}), x^* - z_{k+1}\rangle. \end{aligned}$$

Thus, we can bound the difference of the discrete-time energy function (99) as follows:

$$\begin{aligned} &\mathcal{E}_{k+1} - \mathcal{E}_k \\ &= (1+\mu A_{k+1}) D_h(x^*, z_{k+1}) - (1+\mu A_k) D_h(x^*, z_k) \\ &\quad + A_{k+1}\left(f(x_{k+1}) - f(x^*)\right) - A_k\left(f(x_k) - f(x^*)\right) \\ &= \mu(A_{k+1} - A_k) D_h(x^*, z_{k+1}) \\ &\quad + (A_{k+1} - A_k)\left(f(x_{k+1}) - f(x^*)\right) + A_k\left(f(x_{k+1}) - f(x_k)\right) \\ &\quad + (1+\mu A_k)\left(-h(z_{k+1}) - \langle \nabla h(z_{k+1}), x^* - z_{k+1}\rangle + h(z_k) + \langle \nabla h(z_k), x^* - z_k\rangle\right) \\ &\le \mu(A_{k+1} - A_k) D_h(x^*, x_{k+1}) - \mu(A_{k+1} - A_k)\langle \nabla h(z_{k+1}) - \nabla h(x_{k+1}), x^* - z_{k+1}\rangle \\ &\quad + (A_{k+1} - A_k)\left(f(x_{k+1}) - f(x^*)\right) + A_k\left(f(x_{k+1}) - f(x_k)\right) \\ &\quad + (1+\mu A_k)\left(-h(z_{k+1}) - \langle \nabla h(z_{k+1}), x^* - z_{k+1}\rangle + h(z_k) + \langle \nabla h(z_k), x^* - z_k\rangle\right). \end{aligned}$$

By the (μ-uniform) convexity of f with respect to h, the p-th order 1-uniform convexity of h, and the inequality (14), the following inequalities hold:

$$\begin{aligned} 0 &\ge f(x_{k+1}) - f(x^*) + \langle \nabla f(x_{k+1}), x^* - x_{k+1}\rangle + \mu D_h(x^*, x_{k+1}) \\ 0 &\ge f(x_{k+1}) - f(x_k) + \langle \nabla f(x_{k+1}), x_k - x_{k+1}\rangle \\ 0 &\ge Ms^{\frac{1}{p-1}} \|\nabla f(x_{k+1})\|^{\frac{p}{p-1}} - \langle \nabla f(x_{k+1}), y_k - x_{k+1}\rangle \\ 0 &\ge h(z_k) - h(z_{k+1}) + \langle \nabla h(z_k), z_{k+1} - z_k\rangle + \frac{1}{p}\|z_{k+1} - z_k\|^p. \end{aligned}$$

[22] $D_h(x,y) - D_h(x,z) = -\langle \nabla h(y) - \nabla h(z), x - y\rangle - D_h(y,z)$ (see Wilson et al., 2021), with $x = x^*$, $y = z_{k+1}$, $z = x_{k+1}$.

Taking a weighted sum of these inequalities yields

$$
\begin{aligned}
0 \geq\ & \left(A_{k+1}-A_k\right)\left[f\left(x_{k+1}\right)-f\left(x^*\right)+\left\langle \nabla f\left(x_{k+1}\right), x^*-x_{k+1}\right\rangle+\mu D_h\left(x^*, x_{k+1}\right)\right] \\
& +A_k\left[f\left(x_{k+1}\right)-f\left(x_k\right)+\left\langle \nabla f\left(x_{k+1}\right), x_k-x_{k+1}\right\rangle\right] \\
& +A_{k+1}\left[M s^{\frac{1}{p-1}}\left\|\nabla f\left(x_{k+1}\right)\right\|^{\frac{p}{p-1}}-\left\langle \nabla f\left(x_{k+1}\right), y_k-x_{k+1}\right\rangle\right] \\
& +\left(1+\mu A_k\right)\left[h\left(z_k\right)-h\left(z_{k+1}\right)+\left\langle \nabla h\left(z_k\right), z_{k+1}-z_k\right\rangle+\frac{1}{p}\left\|z_{k+1}-z_k\right\|^p\right] \\
\geq\ & \mathcal{E}_{k+1}-\mathcal{E}_k \\
& -\mu\left(A_{k+1}-A_k\right) D_h\left(x^*, x_{k+1}\right)+\mu\left(A_{k+1}-A_k\right)\left\langle \nabla h\left(z_{k+1}\right)-\nabla h\left(x_{k+1}\right), x^*-z_{k+1}\right\rangle \\
& -\left(A_{k+1}-A_k\right)\left(f\left(x_{k+1}\right)-f\left(x^*\right)\right)-A_k\left(f\left(x_{k+1}\right)-f\left(x_k\right)\right) \\
& -\left(1+\mu A_{k+1}\right)\left(-h\left(z_{k+1}\right)-\left\langle \nabla h\left(z_{k+1}\right), x^*-z_{k+1}\right\rangle+h\left(z_k\right)+\left\langle \nabla h\left(z_k\right), x^*-z_k\right\rangle\right) \\
& +\left(A_{k+1}-A_k\right)\left[f\left(x_{k+1}\right)-f\left(x^*\right)+\left\langle \nabla f\left(x_{k+1}\right), x^*-x_{k+1}\right\rangle+\mu D_h\left(x^*, x_{k+1}\right)\right] \\
& +A_k\left[f\left(x_{k+1}\right)-f\left(x_k\right)+\left\langle \nabla f\left(x_{k+1}\right), x_k-x_{k+1}\right\rangle\right] \\
& +A_{k+1}\left[M s^{\frac{1}{p-1}}\left\|\nabla f\left(x_{k+1}\right)\right\|^{\frac{p}{p-1}}-\left\langle \nabla f\left(x_{k+1}\right), y_k-x_{k+1}\right\rangle\right] \\
& +\left(1+\mu A_k\right)\left[h\left(z_k\right)-h\left(z_{k+1}\right)+\left\langle \nabla h\left(z_k\right), z_{k+1}-z_k\right\rangle+\frac{1}{p}\left\|z_{k+1}-z_k\right\|^p\right] \\
=\ & \mathcal{E}_{k+1}-\mathcal{E}_k \\
& +\left\langle \nabla f\left(x_{k+1}\right),\left(A_{k+1}-A_k\right)\left(x^*-x_{k+1}\right)+A_k\left(x_k-x_{k+1}\right)+A_{k+1}\left(x_{k+1}-y_k\right)\right\rangle \\
& +\left(1+\mu A_k\right)\left\langle \nabla h\left(z_{k+1}\right)-\nabla h\left(z_k\right), x^*-z_{k+1}\right\rangle+\frac{1+\mu A_k}{p}\left\|z_{k+1}-z_k\right\|^p \\
& +\mu\left(A_{k+1}-A_k\right)\left\langle \nabla h\left(z_{k+1}\right)-\nabla h\left(x_{k+1}\right), x^*-z_{k+1}\right\rangle+M A_{k+1} s^{\frac{1}{p-1}}\left\|\nabla f\left(x_{k+1}\right)\right\|^{\frac{p}{p-1}} .
\end{aligned}
$$

Substituting (105), we have

$$
\begin{aligned}
0 \geq\ & \mathcal{E}_{k+1}-\mathcal{E}_k \\
& +\left\langle \nabla f\left(x_{k+1}\right),\left(A_{k+1}-A_k\right)\left(x^*-x_{k+1}\right)+A_k\left(x_k-x_{k+1}\right)+A_{k+1}\left(x_{k+1}-y_k\right)\right\rangle \\
& +\left(A_{k+1}-A_k\right)\left\langle \mu \nabla h\left(x_{k+1}\right)-\mu \nabla h\left(z_{k+1}\right)-\nabla f\left(x_{k+1}\right), x^*-z_{k+1}\right\rangle \\
& +\frac{1+\mu A_k}{p}\left\|z_{k+1}-z_k\right\|^p \\
& +\mu\left(A_{k+1}-A_k\right)\left\langle \nabla h\left(z_{k+1}\right)-\nabla h\left(x_{k+1}\right), x^*-z_{k+1}\right\rangle+M A_{k+1} s^{\frac{1}{p-1}}\left\|\nabla f\left(x_{k+1}\right)\right\|^{\frac{p}{p-1}} \\
=\ & \mathcal{E}_{k+1}-\mathcal{E}_k \\
& +\left\langle \nabla f\left(x_{k+1}\right),\left(A_{k+1}-A_k\right)\left(z_{k+1}-x_{k+1}\right)+A_k\left(x_k-x_{k+1}\right)+A_{k+1}\left(x_{k+1}-y_k\right)\right\rangle \\
& +\frac{1+\mu A_k}{p}\left\|z_{k+1}-z_k\right\|^p+M A_{k+1} s^{\frac{1}{p-1}}\left\|\nabla f\left(x_{k+1}\right)\right\|^{\frac{p}{p-1}} .
\end{aligned}
$$

We also notice that

$$
\begin{aligned}
& \left(A_{k+1}-A_k\right)\left(z_{k+1}-x_{k+1}\right)+A_k\left(x_k-x_{k+1}\right)+A_{k+1}\left(x_{k+1}-y_k\right) \\
& =\left(A_{k+1}-A_k\right) z_{k+1}+A_k x_k-A_{k+1} y_k \\
& =\left(A_{k+1}-A_k\right)\left(z_{k+1}-z_k\right)+\left(A_{k+1}-A_k\right) z_k+A_k x_k-A_{k+1} y_k \\
& =\left(A_{k+1}-A_k\right)\left(z_{k+1}-z_k\right),
\end{aligned}
$$

where the last equality follows from $y_k=x_k+\frac{A_{k+1}-A_k}{A_{k+1}}\left(z_k-x_k\right)$. Therefore,

$$
\begin{aligned}
0 \geq\ & \mathcal{E}_{k+1}-\mathcal{E}_k \\
& +\left(A_{k+1}-A_k\right)\left\langle \nabla f\left(x_{k+1}\right), z_{k+1}-z_k\right\rangle \\
& +\frac{1+\mu A_k}{p}\left\|z_{k+1}-z_k\right\|^p+M A_{k+1} s^{\frac{1}{p-1}}\left\|\nabla f\left(x_{k+1}\right)\right\|^{\frac{p}{p-1}} .
\end{aligned}
$$

Now, we use the Fenchel-Young inequality $\langle s,u\rangle + \frac{1}{p}\|u\|^p \geq -\frac{p-1}{p}\|s\|^{\frac{p}{p-1}}$ (Nesterov, 2008, Lemma 2) with $u = (1+\mu A_k)^{\frac{1}{p}}(z_{k+1}-z_k)$ and $s = (A_{k+1}-A_k)(1+\mu A_k)^{-\frac{1}{p}}\nabla f(x_{k+1})$ to obtain that

$$(A_{k+1}-A_k)\langle \nabla f(x_{k+1}), z_{k+1}-z_k\rangle + \frac{1+\mu A_k}{p}\|z_{k+1}-z_k\|^p \geq -\frac{p-1}{p}(A_{k+1}-A_k)^{\frac{p}{p-1}}(1+\mu A_k)^{-\frac{1}{p-1}}\|\nabla f(x_{k+1})\|^{\frac{p}{p-1}}.$$

Hence, we have

$$\begin{aligned}0 &\geq \mathcal{E}_{k+1}-\mathcal{E}_k \\ &\quad + \left(MA_{k+1}s^{\frac{1}{p-1}} - \frac{p-1}{p}(A_{k+1}-A_k)^{\frac{p}{p-1}}(1+\mu A_k)^{-\frac{1}{p-1}}\right)\|\nabla f(x_{k+1})\|^{\frac{p}{p-1}} \\ &= \mathcal{E}_{k+1}-\mathcal{E}_k \\ &\quad + \left((p-1)p^{\frac{1}{p-1}}C^{\frac{1}{p-1}}A_{k+1}s^{\frac{1}{p-1}} - \frac{p-1}{p}(A_{k+1}-A_k)^{\frac{p}{p-1}}(1+\mu A_k)^{-\frac{1}{p-1}}\right)\|\nabla f(x_{k+1})\|^{\frac{p}{p-1}},\end{aligned}$$

where $C = \frac{1}{p}(\frac{M}{p-1})^{p-1}$. It is easy to see that the sequence (A_k) satisfies

$$(A_{k+1}-A_k)^p - Cp^p s A_{k+1}^{p-1}(1+\mu A_k) \leq 0. \tag{100}$$

Thus, the term

$$\left((p-1)p^{\frac{1}{p-1}}C^{\frac{1}{p-1}}A_{k+1}s^{\frac{1}{p-1}} - \frac{p-1}{p}(A_{k+1}-A_k)^{\frac{p}{p-1}}(1+\mu A_k)^{-\frac{1}{p-1}}\right)\|\nabla f(x_{k+1})\|^{\frac{p}{p-1}}$$

is non-negative. Thus, we have $0 \geq \mathcal{E}_{k+1}-\mathcal{E}_k$ and we conclude that the energy function (99) is non-increasing. Writing $\mathcal{E}_k \leq \mathcal{E}_0$ explicitly, we obtain

$$f(x_k) - f(x^*) \leq \frac{1}{A_k}\left((1+\mu A_0)D_h(x^*,x_0) + A_0(f(x_0)-f(x^*))\right).$$

Now, we show that the unified ATM achieves an $O(1/k^p)$ convergence rate. Let $B_k = \sqrt[p]{A_k}$. Then, the updating rule of A_k implies

$$B_{k+1}-B_k = A_{k+1}^{1/p} - A_k^{1/p} \geq \left(A_k + C^{1/p}ps^{1/p}A_k^{\frac{p-1}{p}}\right)^{1/p} - A_k^{1/p} = \left(B_k^p + C^{1/p}ps^{1/p}B_k^{p-1}\right)^{1/p} - B_k.$$

Because $\lim_{k\to\infty}B_k = \infty$ and we have $\lim_{X\to\infty}\left\{\left(X^p + p\alpha X^{p-1}\right)^{1/p} - X\right\} = \alpha$ for all $\alpha > 0$, we have $\liminf_{k\to\infty}\{B_{k+1}-B_k\} > 0$. Thus, we hve $A_k = \Omega(k^p)$, which implies that the unified ATM achieves an $O(1/k^p)$ convergence rate. Next, it is easy to show that the unified ATM achieves an $O((1+p\sqrt[p]{C\mu s})^{-k})$ convergence rate because

$$A_{k+1} = A_k + p\sqrt[p]{CsA_k^{p-1}(1+\mu A_k)} \geq A_k + p\sqrt[p]{CsA_k^{p-1}\cdot\mu A_k} = \left(1+p\sqrt[p]{C\mu s}\right)A_k.$$

This completes the proof of Theorem 5.3.

Remark E.1. We show that $(1+p\sqrt[p]{C\mu s})^{-k} \leq \exp(-\frac{1}{9}\sqrt[p]{\mu s}k)$ holds when $N = \sqrt{2}$ and $M = 1/3$. Because $C = \frac{1}{p}\left(\frac{M}{p-1}\right)^{p-1}$, we have $C^{1/p}p = \left(\frac{Mp}{p-1}\right)^{1-\frac{1}{p}} \geq M^{1-\frac{1}{p}} \geq M = 1/3$. Thus, we have

$$(1+C^{1/p}p\mu^{1/p}s^{1/p})^{-k} \leq \left(1+\frac{1}{3}(\mu s)^{1/p}\right)^{-k} = \exp\left(-k\log\left(1+\frac{1}{3}(\mu s)^{1/p}\right)\right).$$

Using the fact that the inequality $\log(1+\frac{x}{3}) \geq x/9$ holds for all $x \in [0,1]$, we obtain

$$(1+C^{1/p}p\mu^{1/p}s^{1/p})^{-k} \leq e^{-\frac{1}{9}\mu^{1/p}s^{1/p}k}.$$

Remark E.2. When $\mu = 0$, the unified ATM ODE (12) recovers the ATM-C ODE given in (Wibisono et al., 2016):

$$\begin{aligned} \dot{X} &= \frac{p}{t}(Z - X) \\ \frac{d}{dt}\nabla h(Z) &= -Cpt^{p-1}\nabla f(X). \end{aligned} \tag{101}$$

Moreover, the updating rules for unified ATM can be written as

$$\begin{aligned} y_k &= x_k + \frac{A_{k+1} - A_k}{A_{k+1}}(z_k - x_k) \\ x_{k+1} &= G_{p,s,N}(y_k) \\ z_{k+1} &= \arg\min_z \left\{ (A_{k+1} - A_k)\langle \nabla f(x_{k+1}), z\rangle + D_h(z, z_k) \right\}. \end{aligned} \tag{102}$$

To exactly recover ATM-C, we modify the sequence (A_k) as $A_k := Csk(k+1)\cdots(k+p-1)$, which is simpler and also satisfies $A_k = \Omega(k^p)$. Then, this sequence satisfies (100) because

$$\begin{aligned} C^{\frac{1}{p-1}} p^{\frac{p}{p-1}} s^{\frac{1}{p-1}} A_{k+1} &= C^{\frac{1}{p-1}} p^{\frac{p}{p-1}} s^{\frac{1}{p-1}} \cdot Cs(k+1)\cdots(k+p) \\ &\geq C^{\frac{1}{p-1}} p^{\frac{p}{p-1}} s^{\frac{1}{p-1}} \cdot Cs(k+1)^{\frac{p}{p-1}} \cdots (k+p-1)^{\frac{p}{p-1}} \\ &= (A_{k+1} - A_k)^{\frac{p}{p-1}}. \end{aligned}$$

Thus, this modification does not affect the validity of the convergence rate (15). In this case, Algorithm 2 and its convergence rate recover ATM-C and its convergence rate (Wibisono et al., 2016, Equation 20).

E.4. Unified ATM Converges to Unified ATM ODE

Let $\mathbf{t}_k := \sqrt[p]{A_k/C}$ if $\mu = 0$ and $\mathbf{t}_k := \sinh_p^{-1}(\sqrt[p]{\mu A_k})/\sqrt[p]{C\mu}$ if $\mu > 0$. We first that the timesteps are asymptotically equivalent to $s^{1/p}$ as $s \to 0$ in the sense that

$$\lim_{s\to 0} \frac{\mathbf{t}_{\mathbf{k}(t)+1} - t}{s^{1/p}} = 1 \quad \forall t \in (0, \infty), \tag{103}$$

where $\mathbf{k}$ is the inverse of $\mathbf{t}$. It is easy to check that the function $A(t)$ defined in (104) satisfies

$$\dot{A}(t) = C^{1/p} p \mu^{\frac{1-p}{p}} \sinh_p^{p-1}\left(C^{1/p}\mu^{1/p}t\right)\cosh_p\left(C^{1/p}\mu^{1/p}t\right) = C^{1/p} p A(t)^{\frac{p-1}{p}}(1 + \mu A(t))^{\frac{1}{p}}$$

and that the sequence $(\mathbf{t}_k)$ satisfies

$$\frac{A(\mathbf{t}_{k+1}) - A(\mathbf{t}_k)}{s^{1/p}} - C^{1/p} p A(\mathbf{t}_{k+1})^{\frac{p-1}{p}}(1 + \mu A(\mathbf{t}_k))^{\frac{1}{p}} = 0.$$

Now, substituting $k = \mathbf{k}(t)$ into the above equality and taking the limit $s \to 0$, we have $\lim_{s\to 0} \frac{\mathbf{t}_{\mathbf{k}(t)+1} - t}{s^{1/p}} = 1$.

Now, using (103), we show that the unified ATM converges to the unified ATM ODE (12) under the identifications $x_k \leftrightarrow X(\mathbf{t}_k)$ and $z_k \leftrightarrow Z(\mathbf{t}_k)$. For convenience, we assume that $\mu > 0$ (the case $\mu = 0$ can be handled easily). Define a function $A : [0, \infty) \to \mathbb{R}$ as

$$A(t) = Ct^p \operatorname{sinhc}_p^p\left(C^{1/p}\mu^{1/p}t\right) = \frac{1}{\mu}\sinh_p^p\left(C^{1/p}\mu^{1/p}t\right), \tag{104}$$

so that $A_k = A(\mathbf{t}_k)$. Then, we have

$$\begin{aligned} \dot{X}(t) &= \lim_{s\to 0} \frac{x_{\mathbf{k}(t)+1} - x_{\mathbf{k}(t)}}{\mathbf{t}_{\mathbf{k}(t)+1} - t} \\ &= \lim_{s\to 0} \frac{x_{\mathbf{k}(t)+1} - x_{\mathbf{k}(t)}}{s^{1/p}} \end{aligned}$$

$$
\begin{aligned}
&= \lim_{s\to 0} \frac{y_{\mathbf{k}(t)} - x_{\mathbf{k}(t)}}{s^{1/p}} \\
&= \lim_{s\to 0} \frac{A_{\mathbf{k}(t)+1} - A_{\mathbf{k}(t)}}{s^{1/p} A_{\mathbf{k}(t)+1}} \left(z_{\mathbf{k}(t)} - x_{\mathbf{k}(t)}\right) \\
&= \lim_{s\to 0} \frac{A\left(\mathbf{t}_{\mathbf{k}(t)+1}\right) - A(t)}{s^{1/p} A\left(\mathbf{t}_{\mathbf{k}(t)+1}\right)} \left(Z(t) - X(t)\right) \\
&= \frac{\dot{A}(t)}{A(t)} \left(Z(t) - X(t)\right) \\
&= pC^{1/p}\mu^{1/p} \coth_p\left(C^{1/p}\mu^{1/p} t\right) \left(Z(t) - X(t)\right),
\end{aligned}
$$

where we used $\|x_{k+1} - y_k\| = o(s^{1/p})$ (see Wibisono et al., 2016, Lemma 2.2) for the third equality.

By the first-order optimality condition, the updating rule for z_k in the unified ATM is equivalent to

$$
\nabla h\left(z_{k+1}\right) - \nabla h\left(z_k\right) = \frac{A_{k+1} - A_k}{1 + \mu A_k} \left(\mu \nabla h\left(x_{k+1}\right) - \mu \nabla h\left(z_{k+1}\right) - \nabla f\left(x_{k+1}\right)\right). \tag{105}
$$

Thus, we have

$$
\begin{aligned}
\frac{d}{dt}\nabla h(Z(t)) &= \lim_{s\to 0} \frac{\nabla h\left(z_{\mathbf{k}(t)+1}\right) - \nabla h\left(z_{\mathbf{k}(t)}\right)}{\mathbf{t}_{\mathbf{k}(t)+1} - t} \\
&= \lim_{s\to 0} \frac{\nabla h\left(z_{\mathbf{k}(t)+1}\right) - \nabla h\left(z_{\mathbf{k}(t)}\right)}{s^{1/p}} \\
&= \lim_{s\to 0} \frac{A_{\mathbf{k}(t)+1} - A_{\mathbf{k}(t)}}{s^{1/p}\left(1 + \mu A_{\mathbf{k}(t)}\right)} \left(\mu \nabla h\left(x_{\mathbf{k}(t)+1}\right) - \mu \nabla h\left(z_{\mathbf{k}(t)+1}\right) - \nabla f\left(x_{\mathbf{k}(t)+1}\right)\right) \\
&= \lim_{s\to 0} \frac{A\left(\mathbf{t}_{\mathbf{k}(t)+1}\right) - A(t)}{s^{1/p}\left(1 + \mu A(t)\right)} \left(\mu \nabla h\left(X(t)\right) - \mu \nabla h\left(X(t)\right) - \nabla f\left(X(t)\right)\right) \\
&= \frac{\dot{A}(t)}{1 + \mu A(t)} \left(\mu \nabla h\left(X(t)\right) - \mu \nabla h\left(X(t)\right) - \nabla f\left(X(t)\right)\right) \\
&= \frac{C^{1/p} p}{\mu^{(p-1)/p}} \tanh_p^{p-1}\left(C^{1/p}\mu^{1/p} t\right) \left(\mu \nabla h\left(X(t)\right) - \mu \nabla h\left(X(t)\right) - \nabla f\left(X(t)\right)\right).
\end{aligned}
$$

Thus, we conclude that the unified ATM converges to the unified ATM ODE (12).

E.5. ATM-SC ODE: Asymptotic Limit of Unified ATM ODE

In this subsection, we investigate the asymptotic limit of the unified ATM ODE (12) with $\mu > 0$. Because the unified ATM ODE is the unified Bregman Lagrangian flow (5) with $\alpha(t) = \log(\frac{p}{t}\operatorname{cothc}_p(\sqrt[p]{C\mu}t))$ and $\beta(t) = \log(Ct^p \operatorname{sinhc}_p^p(\sqrt[p]{C\mu}t))$, its asymptotic limit is the system (7) with $\alpha(\infty) = \log(C^{1/p}p\mu^{1/p})$ and $\dot{\beta}(\infty) = C^{1/p}p\mu^{1/p}$:

$$
\begin{aligned}
\dot{X} &= C^{1/p}p\mu^{1/p}(Z - X) \\
\frac{d}{dt}\nabla h(Z) &= C^{1/p}p\mu^{1/p}\left(\nabla h(X) - \nabla h(Z) - \frac{1}{\mu}\nabla f(X)\right),
\end{aligned}
$$

which we call ATM-SC ODE. Because this system is the second Bregman Lagrangian flow (27) with $\alpha_{2\mathrm{nd}}(t) :\equiv \alpha(\infty) = \log(C^{1/p}p\mu^{1/p})$ and $\beta_{2\mathrm{nd}}(t) := \dot{\beta}(\infty)t = C^{1/p}p\mu^{1/p}t$, it achieves the following convergence rate (see also Appendix C.5):

$$
f(X(t)) - f(x^*) \le O\left(e^{-\beta_{2\mathrm{nd}}(t)}\right) = O\left(e^{-C^{1/p}p\mu^{1/p}t}\right).
$$

E.6. ATM-SC: Asymptotic Limit of Unified ATM

In this subsection, we investigate the asymptotic limit of the unified ATM (Algorithm 2) with $\mu > 0$. Because we have $\lim_{k\to\infty} A_k = \infty$ and

$$\lim_{k\to\infty} \frac{A_{k+1}}{A_k} = \lim_{k\to\infty} \frac{A_k + p\sqrt[p]{CsA_k^{p-1}(1+\mu A_k)}}{A_k} = 1 + p\sqrt[p]{C\mu s},$$

we can compute the momentum coefficients in the unified ATM (Algorithm 2) as

$$\lim_{k\to\infty} \frac{A_{k+1} - A_k}{A_{k+1}} = \frac{p\sqrt[p]{C\mu s}}{1 + p\sqrt[p]{C\mu s}}$$
$$\lim_{k\to\infty} \frac{A_{k+1} - A_k}{1 + \mu A_k} = \frac{p\sqrt[p]{C\mu s}}{\mu}.$$

Replacing the momentum coefficients $\frac{A_{k+1}-A_k}{A_{k+1}}$ and $\frac{A_{k+1}-A_k}{1+\mu A_k}$ with their limits in the unified ATM algorithm, we yield the following time-invariant algorithm, which we call ATM-SC:

$$\begin{aligned} y_k &= x_k + \frac{p\sqrt[p]{C\mu s}}{1 + p\sqrt[p]{C\mu s}}(z_k - x_k) \\ x_{k+1} &= G_{p,s,N}(y_k) \\ z_{k+1} &= \arg\min_z \left\{ \frac{p\sqrt[p]{C\mu s}}{\mu}(\langle \nabla f(x_{k+1}), z\rangle + \mu D_h(z, x_{k+1})) + D_h(z, z_k) \right\}. \end{aligned}$$

Following the argument in Appendix D.7, we can show that ATM-SC (Algorithm 3) is the asymptotic limit of the unified ATM, in the sense that the output of the unified ATM converges to the one of ATM-SC as $k_0 \to \infty$ (the proof is similar, so we omit it). The convergence rate of this algorithm is addressed in the following theorem.

Theorem E.3. *The iterates of ATM-SC (Algorithm 3) with $s \le (p-1)!/L$ satisfy*

$$f(x_k) - f(x^*) \le O((1 + p\sqrt[p]{C\mu s})^{-k}).$$

Proof. For convenience, we define a sequence (A_k) as

$$A_k = \left(1 + p\sqrt[p]{Cs\mu}\right)^k.$$

Then, the updating rules in ATM-SC can be written as

$$\begin{aligned} y_k &= x_k + \frac{A_{k+1} - A_k}{A_{k+1}}(z_k - x_k) \\ x_{k+1} &= G_{p,s,N}(y_k) \\ z_{k+1} &= \arg\min_z \left\{ \frac{A_{k+1} - A_k}{\mu A_k}(\langle \nabla f(x_{k+1}), z\rangle + \mu D_h(z, x_{k+1})) + D_h(z, z_k) \right\}. \end{aligned}$$

Define a discrete-time energy function $\mathcal{E}_k$ as

$$\mathcal{E}_k = A_k \left(f\left(x_k\right) - f\left(x^*\right) + \mu D_h\left(x^*, z_k\right)\right).$$

By the Bregman three-point identity[23] and the non-negativity of Bregman divergence, we have

$$\begin{aligned} D_h\left(x^*, z_{k+1}\right) &= D_h\left(x^*, x_{k+1}\right) - \left\langle \nabla h\left(z_{k+1}\right) - \nabla h\left(x_{k+1}\right), x^* - z_{k+1}\right\rangle - D_h\left(z_{k+1}, x_{k+1}\right) \\ &\le D_h\left(x^*, x_{k+1}\right) - \left\langle \nabla h\left(z_{k+1}\right) - \nabla h\left(x_{k+1}\right), x^* - z_{k+1}\right\rangle. \end{aligned}$$

[23] $D_h(x,y) - D_h(x,z) = -\langle \nabla h(y) - \nabla h(z), x - y\rangle - D_h(y,z)$ (see Wilson et al., 2021), with $x = x^*$, $y = z_{k+1}$, $z = x_{k+1}$.

Thus, we can bound the difference of the discrete-time energy function as follows:

$$
\begin{aligned}
&\mathcal{E}_{k+1}-\mathcal{E}_k\\
&=\mu A_{k+1}D_h\left(x^*,z_{k+1}\right)-\mu A_kD_h\left(x^*,z_k\right)\\
&\quad+A_{k+1}\left(f\left(x_{k+1}\right)-f\left(x^*\right)\right)-A_k\left(f\left(x_k\right)-f\left(x^*\right)\right)\\
&=\mu\left(A_{k+1}-A_k\right)D_h\left(x^*,z_{k+1}\right)\\
&\quad+\left(A_{k+1}-A_k\right)\left(f\left(x_{k+1}\right)-f\left(x^*\right)\right)+A_k\left(f\left(x_{k+1}\right)-f\left(x_k\right)\right)\\
&\quad+\mu A_k\left(-h\left(z_{k+1}\right)-\left\langle\nabla h\left(z_{k+1}\right),x^*-z_{k+1}\right\rangle+h\left(z_k\right)+\left\langle\nabla h\left(z_k\right),x^*-z_k\right\rangle\right)\\
&\leq\mu\left(A_{k+1}-A_k\right)D_h\left(x^*,x_{k+1}\right)-\mu\left(A_{k+1}-A_k\right)\left\langle\nabla h\left(z_{k+1}\right)-\nabla h\left(x_{k+1}\right),x^*-z_{k+1}\right\rangle\\
&\quad+\left(A_{k+1}-A_k\right)\left(f\left(x_{k+1}\right)-f\left(x^*\right)\right)+A_k\left(f\left(x_{k+1}\right)-f\left(x_k\right)\right)\\
&\quad+\mu A_k\left(-h\left(z_{k+1}\right)-\left\langle\nabla h\left(z_{k+1}\right),x^*-z_{k+1}\right\rangle+h\left(z_k\right)+\left\langle\nabla h\left(z_k\right),x^*-z_k\right\rangle\right).
\end{aligned}
$$

By the μ-uniform convexity of f with respect to h, the p-th order 1-uniform convexity of h, and the property (14) of the higher-order gradient update operator $G_{p,M}$, the following inequalities hold:

$$
\begin{aligned}
&0\geq f\left(x_{k+1}\right)-f\left(x^*\right)+\left\langle\nabla f\left(x_{k+1}\right),x^*-x_{k+1}\right\rangle+\mu D_h\left(x^*,x_{k+1}\right)\\
&0\geq f\left(x_{k+1}\right)-f\left(x_k\right)+\left\langle\nabla f\left(x_{k+1}\right),x_k-x_{k+1}\right\rangle\\
&0\geq Ms^{\frac{1}{p-1}}\left\|\nabla f\left(x_{k+1}\right)\right\|^{\frac{p}{p-1}}-\left\langle\nabla f\left(x_{k+1}\right),y_k-x_{k+1}\right\rangle\\
&0\geq h\left(z_k\right)-h\left(z_{k+1}\right)+\left\langle\nabla h\left(z_k\right),z_{k+1}-z_k\right\rangle+\frac{1}{p}\left\|z_{k+1}-z_k\right\|^p.
\end{aligned}
$$

Taking a weighted sum of these inequalities yields

$$
\begin{aligned}
0\geq&\left(A_{k+1}-A_k\right)\left[f\left(x_{k+1}\right)-f\left(x^*\right)+\left\langle\nabla f\left(x_{k+1}\right),x^*-x_{k+1}\right\rangle+\mu D_h\left(x^*,x_{k+1}\right)\right]\\
&+A_k\left[f\left(x_{k+1}\right)-f\left(x_k\right)+\left\langle\nabla f\left(x_{k+1}\right),x_k-x_{k+1}\right\rangle\right]\\
&+A_{k+1}\left[Ms^{\frac{1}{p-1}}\left\|\nabla f\left(x_{k+1}\right)\right\|^{\frac{p}{p-1}}-\left\langle\nabla f\left(x_{k+1}\right),y_k-x_{k+1}\right\rangle\right]\\
&+\mu A_k\left[h\left(z_k\right)-h\left(z_{k+1}\right)+\left\langle\nabla h\left(z_k\right),z_{k+1}-z_k\right\rangle+\frac{1}{p}\left\|z_{k+1}-z_k\right\|^p\right]\\
\geq&\,\mathcal{E}_{k+1}-\mathcal{E}_k\\
&-\mu\left(A_{k+1}-A_k\right)D_h\left(x^*,x_{k+1}\right)+\mu\left(A_{k+1}-A_k\right)\left\langle\nabla h\left(z_{k+1}\right)-\nabla h\left(x_{k+1}\right),x^*-z_{k+1}\right\rangle\\
&-\left(A_{k+1}-A_k\right)\left(f\left(x_{k+1}\right)-f\left(x^*\right)\right)-A_k\left(f\left(x_{k+1}\right)-f\left(x_k\right)\right)\\
&-\mu A_{k+1}\left(-h\left(z_{k+1}\right)-\left\langle\nabla h\left(z_{k+1}\right),x^*-z_{k+1}\right\rangle+h\left(z_k\right)+\left\langle\nabla h\left(z_k\right),x^*-z_k\right\rangle\right)\\
&+\left(A_{k+1}-A_k\right)\left[f\left(x_{k+1}\right)-f\left(x^*\right)+\left\langle\nabla f\left(x_{k+1}\right),x^*-x_{k+1}\right\rangle+\mu D_h\left(x^*,x_{k+1}\right)\right]\\
&+A_k\left[f\left(x_{k+1}\right)-f\left(x_k\right)+\left\langle\nabla f\left(x_{k+1}\right),x_k-x_{k+1}\right\rangle\right]\\
&+A_{k+1}\left[Ms^{\frac{1}{p-1}}\left\|\nabla f\left(x_{k+1}\right)\right\|^{\frac{p}{p-1}}-\left\langle\nabla f\left(x_{k+1}\right),y_k-x_{k+1}\right\rangle\right]\\
&+\mu A_k\left[h\left(z_k\right)-h\left(z_{k+1}\right)+\left\langle\nabla h\left(z_k\right),z_{k+1}-z_k\right\rangle+\frac{1}{p}\left\|z_{k+1}-z_k\right\|^p\right]\\
=&\,\mathcal{E}_{k+1}-\mathcal{E}_k\\
&+\left\langle\nabla f\left(x_{k+1}\right),\left(A_{k+1}-A_k\right)\left(x^*-x_{k+1}\right)+A_k\left(x_k-x_{k+1}\right)+A_{k+1}\left(x_{k+1}-y_k\right)\right\rangle\\
&+\mu A_k\left\langle\nabla h\left(z_{k+1}\right)-\nabla h\left(z_k\right),x^*-z_{k+1}\right\rangle+\frac{\mu A_k}{p}\left\|z_{k+1}-z_k\right\|^p\\
&+\mu\left(A_{k+1}-A_k\right)\left\langle\nabla h\left(z_{k+1}\right)-\nabla h\left(x_{k+1}\right),x^*-z_{k+1}\right\rangle+MA_{k+1}s^{\frac{1}{p-1}}\left\|\nabla f\left(x_{k+1}\right)\right\|^{\frac{p}{p-1}}.
\end{aligned}
$$

By the first-order optimality condition, the updating rule for z_k is equivalent to

$$
\nabla h\left(z_{k+1}\right)-\nabla h\left(z_k\right)=\frac{A_{k+1}-A_k}{\mu A_k}\left(\mu\nabla h\left(x_{k+1}\right)-\mu\nabla h\left(z_{k+1}\right)-\nabla f\left(x_{k+1}\right)\right). \tag{106}
$$

Substituting this, we have

$$\begin{aligned}
0 \geq\ & \mathcal{E}_{k+1} - \mathcal{E}_k \\
& + \left\langle \nabla f\left(x_{k+1}\right), \left(A_{k+1} - A_k\right)\left(x^* - x_{k+1}\right) + A_k\left(x_k - x_{k+1}\right) + A_{k+1}\left(x_{k+1} - y_k\right)\right\rangle \\
& + \left(A_{k+1} - A_k\right)\left\langle \mu\nabla h\left(x_{k+1}\right) - \mu\nabla h\left(z_{k+1}\right) - \nabla f\left(x_{k+1}\right), x^* - z_{k+1}\right\rangle \\
& + \frac{\mu A_k}{p}\left\|z_{k+1} - z_k\right\|^p \\
& + \mu\left(A_{k+1} - A_k\right)\left\langle \nabla h\left(z_{k+1}\right) - \nabla h\left(x_{k+1}\right), x^* - z_{k+1}\right\rangle + MA_{k+1}s^{\frac{1}{p-1}}\left\|\nabla f\left(x_{k+1}\right)\right\|^{\frac{p}{p-1}} \\
=\ & \mathcal{E}_{k+1} - \mathcal{E}_k \\
& + \left\langle \nabla f\left(x_{k+1}\right), \left(A_{k+1} - A_k\right)\left(z_{k+1} - x_{k+1}\right) + A_k\left(x_k - x_{k+1}\right) + A_{k+1}\left(x_{k+1} - y_k\right)\right\rangle \\
& + \frac{\mu A_k}{p}\left\|z_{k+1} - z_k\right\|^p + MA_{k+1}s^{\frac{1}{p-1}}\left\|\nabla f\left(x_{k+1}\right)\right\|^{\frac{p}{p-1}}.
\end{aligned}$$

We also notice that

$$\begin{aligned}
& \left(A_{k+1} - A_k\right)\left(z_{k+1} - x_{k+1}\right) + A_k\left(x_k - x_{k+1}\right) + A_{k+1}\left(x_{k+1} - y_k\right) \\
&= \left(A_{k+1} - A_k\right)z_{k+1} + A_k x_k - A_{k+1}y_k \\
&= \left(A_{k+1} - A_k\right)\left(z_{k+1} - z_k\right) + \left(A_{k+1} - A_k\right)z_k + A_k x_k - A_{k+1}y_k \\
&= \left(A_{k+1} - A_k\right)\left(z_{k+1} - z_k\right),
\end{aligned}$$

where the last equality follows from $y_k = x_k + \frac{A_{k+1} - A_k}{A_{k+1}}\left(z_k - x_k\right)$. Therefore,

$$\begin{aligned}
0 \geq\ & \mathcal{E}_{k+1} - \mathcal{E}_k \\
& + \left(A_{k+1} - A_k\right)\left\langle \nabla f\left(x_{k+1}\right), z_{k+1} - z_k\right\rangle \\
& + \frac{\mu A_k}{p}\left\|z_{k+1} - z_k\right\|^p + MA_{k+1}s^{\frac{1}{p-1}}\left\|\nabla f\left(x_{k+1}\right)\right\|^{\frac{p}{p-1}}.
\end{aligned}$$

Now, we use the Fenchel-Young inequality $\langle s, u\rangle + \frac{1}{p}\|u\|^p \geq -\frac{p-1}{p}\|s\|^{\frac{p}{p-1}}$ with $u = \mu^{1/p}A_k^{1/p}\left(z_{k+1} - z_k\right)$ and $s = \left(A_{k+1} - A_k\right)\mu^{-1/p}A_k^{-1/p}\nabla f\left(x_{k+1}\right)$ to obtain that

$$\begin{aligned}
& \left(A_{k+1} - A_k\right)\left\langle \nabla f\left(x_{k+1}\right), z_{k+1} - z_k\right\rangle + \frac{\mu A_k}{p}\left\|z_{k+1} - z_k\right\|^p \\
& \qquad\qquad \geq -\frac{p-1}{p}\left(A_{k+1} - A_k\right)^{\frac{p}{p-1}}\mu^{-\frac{1}{p-1}}A_k^{-\frac{1}{p-1}}\left\|\nabla f\left(x_{k+1}\right)\right\|^{\frac{p}{p-1}}.
\end{aligned}$$

Hence, we have

$$\begin{aligned}
0 \geq\ & \mathcal{E}_{k+1} - \mathcal{E}_k \\
& + \left(MA_{k+1}s^{\frac{1}{p-1}} - \frac{p-1}{p}\left(A_{k+1} - A_k\right)^{\frac{p}{p-1}}\left(\mu A_k\right)^{-\frac{1}{p-1}}\right)\left\|\nabla f\left(x_{k+1}\right)\right\|^{\frac{p}{p-1}} \\
=\ & \mathcal{E}_{k+1} - \mathcal{E}_k \\
& + \left((p-1)p^{\frac{1}{p-1}}C^{\frac{1}{p-1}}A_{k+1}s^{\frac{1}{p-1}} - \frac{p-1}{p}\left(A_{k+1} - A_k\right)^{\frac{p}{p-1}}\left(\mu A_k\right)^{-\frac{1}{p-1}}\right)\left\|\nabla f\left(x_{k+1}\right)\right\|^{\frac{p}{p-1}},
\end{aligned}$$

where $C = \frac{1}{p}\left(\frac{M}{p-1}\right)^{p-1}$. It is easy to see that the sequence (A_k) satisfies

$$\left(A_{k+1} - A_k\right)^p - Cp^p\mu s A_k A_{k+1}^{p-1} \leq 0. \tag{107}$$

Thus, the term

$$\left((p-1)p^{\frac{1}{p-1}}C^{\frac{1}{p-1}}A_{k+1}s^{\frac{1}{p-1}} - \frac{p-1}{p}\left(A_{k+1} - A_k\right)^{\frac{p}{p-1}}\left(\mu A_k\right)^{-\frac{1}{p-1}}\right)\left\|\nabla f\left(x_{k+1}\right)\right\|^{\frac{p}{p-1}}$$

is non-negative. Thus, we conclude that the energy function (99) is non-increasing. Writing $\mathcal{E}_k \leq \mathcal{E}_0$ explicitly, we obtain

$$f(x_k) - f(x^*) \leq \frac{A_0}{A_k}\left(f(x_0) - f(x^*) + \mu D_h(x^*, x_0)\right),$$

which implies an $O((1 + p\sqrt[p]{C\mu s})^{-k})$ convergence rate. □

F. ODE Model for Minimizing the Gradient Norm of Strongly Convex Functions

F.1. Review: OGM, OGM-G, and Their Limiting ODEs

We review OGM (Kim & Fessler, 2016), an algorithm for reducing the function value accuracy $f(x_N) - f(x^*)$, and OGM-G (Kim & Fessler, 2021), an algorithm for reducing the squared gradient norm $\|\nabla f(x_N)\|^2$. Given the number N of total iterations, define a sequence $(\theta_k)_{k=0}^N$ as

$$\theta_k = \begin{cases} 1 & \text{if } k = 0 \\ \frac{1+\sqrt{4\theta_{k-1}^2+1}}{2} & \text{if } 1 \leq k \leq N-1 \\ \frac{1+\sqrt{8\theta_{k-1}^2+1}}{2} & \text{if } k = N. \end{cases} \tag{108}$$

Then, OGM is equivalent to the fixed-step first-order scheme (2) with (h_{ij}^{F}), and OGM-G is equivalent to the fixed-step first-order scheme (2) with (h_{ij}^{G}), where the entries of (h_{ij}^{F}) and (h_{ij}^{G}) are defined as

$$\begin{aligned} h_{ij}^{\mathrm{F}} &= \begin{cases} \frac{\theta_i - 1}{\theta_{i+1}} h_{i-1,j} & \text{if } j = 0, \ldots, i-2, \\ \frac{\theta_i - 1}{\theta_{i+1}}\left(h_{i-1,i-1} - 1\right) & \text{if } j = i-1, \\ 1 + \frac{2\theta_i - 1}{\theta_{i+1}} & \text{if } j = i, \end{cases} \\ h_{ij}^{\mathrm{G}} &= \begin{cases} \frac{\theta_{N-i-1} - 1}{\theta_{N-i}} h_{i,j+1} & \text{if } j = 0, \ldots, i-2, \\ \frac{\theta_{N-i-1} - 1}{\theta_{N-i}}\left(h_{i,i} - 1\right) & \text{if } j = i-1, \\ 1 + \frac{2\theta_{N-i-1} - 1}{\theta_{N-i}} & \text{if } j = i. \end{cases} \end{aligned} \tag{109}$$

Suh et al. (2022) showed that OGM-G converges to OGM-G ODE as $s \to 0$, under the identifications $x_k \leftrightarrow X(k\sqrt{s})$ and $N \leftrightarrow T/\sqrt{s}$. Next, we provide a simple argument to show that OGM converges to OGM ODE. For the sequence θ_k defined in (108), Su et al. (2016) showed that the algorithm

$$\begin{aligned} y_k &= \left(1 - \frac{1}{\theta_k}\right) x_k + \frac{1}{\theta_k} z_k \\ x_{k+1} &= y_k - s\nabla f(y_k) \\ z_{k+1} &= z_k - s\theta_k \nabla f(y_k) \end{aligned} \tag{110}$$

converges to AGM-C ODE as $s \to 0$ (see Su et al., 2016, Section 2). Because $\|x_{k+1} - y_k\| = o(\sqrt{s})$, we can ignore the gradient descent step $x_{k+1} = y_k - s\nabla f(y_k)$ in both the algorithm (110) and OGM when dealing with their limiting ODEs. Thus, ignoring the gradient descent step, we can see that applying OGM to the objective function f is equivalent to applying the algorithm (110) to the objective function $2f$. Thus, the limiting ODE of OGM is given by

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

which is OGM ODE.

The differential kernel for OGM ODE can be obtained by substituting $b(t) = 3/t$ and $c(t) = 1$ into (60):

$$H(t,\tau) = 2e^{-\int_\tau^t \frac{3}{s}\,ds} = 2e^{-3\log t + 3\log\tau} = \frac{2\tau^3}{t^3}.$$

The differential kernel for OGM-G ODE can be obtained by substituting $b(t) = 3/(T-t)$ and $c(t) = 1$ into (60):

$$H(t,\tau) = 2e^{-\int_\tau^t \frac{3}{T-s}\,ds} = 2e^{3\log(T-t) - 3\log(T-\tau)} = \frac{2(T-t)^3}{(T-\tau)^3}.$$

F.2. Anti-Transpose Relationship Between OGM and OGM-G

Kim & Fessler (2021, Proposition 6.2) observed the following relationship between the matrices (h_{ij}^{F}) and (h_{ij}^{G}) in (109):

$$h_{ij}^{\mathrm{F}} = h_{N-1-j,N-1-i}^{\mathrm{G}} \quad \forall i \text{ and } j. \tag{111}$$

When the condition (111) holds, we say there is an *anti-transpose* relationship between (h_{ij}^{F}) and (h_{ij}^{G}) because the matrix (h_{ij}^{F}) can be obtained by reflecting (h_{ij}^{G}) about its anti-diagonal and vice versa. Now, it is straightforward to see that the anti-transpose relationship (111) between the two matrices is transferred to the anti-transpose relationship (16) between the differential kernels of OGM ODE and OGM-G ODE as $s \to 0$:

$$H^{\mathrm{F}}(t,\tau) = \lim_{s\to 0} h^{\mathrm{F}}_{\frac{t}{\sqrt{s}},\frac{\tau}{\sqrt{s}}} = \lim_{s\to 0} h^{\mathrm{G}}_{(N-1)-\frac{\tau}{\sqrt{s}},(N-1)-\frac{t}{\sqrt{s}}} = H^{\mathrm{G}}(T-\tau, T-t),$$

where we identify $T = N\sqrt{s}$. To summarize, the relationships between OGM, OGM-G, and their limiting ODEs are illustrated in Figure 6.

$$\begin{array}{ccc}
\underset{\text{(OGM)}}{y_{i+1} = y_i - s\sum_{j=0}^{i} h_{ij}^{\mathrm{F}} \nabla f(y_j)} & \xrightarrow{\text{limiting}} & \underset{\text{(OGM ODE)}}{\dot{X}(t) = -\int_0^t H^{\mathrm{F}}(t,\tau)\nabla f(X(\tau))\, d\tau} \\
h_{i,j}^{\mathrm{F}} = h_{N-1-j,N-1-i}^{\mathrm{G}} \Big\updownarrow & & \Big\updownarrow H^{\mathrm{F}}(t,\tau) = H^{\mathrm{G}}(T-\tau,T-t) \\
\underset{\text{(OGM-G)}}{y_{i+1} = y_i - s\sum_{j=0}^{i} h_{ij}^{\mathrm{G}} \nabla f(y_j)} & \xrightarrow[\text{limiting}]{} & \underset{\text{(OGM-G ODE)}}{\dot{X}(t) = -\int_0^t H^{\mathrm{G}}(t,\tau)\nabla f(X(\tau))\, d\tau}
\end{array}$$

Figure 6. Relationships between OGM (reducing $f(x_k) - f(x^*)$), OGM-G (reducing $\|\nabla f(x_k)\|$), and their limiting ODEs.

F.3. Anti-Transpose Relationship Between Unified AGM ODE and Unified AGM-G ODE

For convenience, assume $\mu > 0$. Substituting $b(t) = \frac{\sqrt{\mu}}{2}\tanh(\frac{\sqrt{\mu}}{2}(T-t)) + \frac{3\sqrt{\mu}}{2}\coth(\frac{\sqrt{\mu}}{2}(T-t))$ and $c(t) = 0$ into (60), we yield the following differential kernel corresponding to the unified AGM-G ODE (17):

$$\begin{aligned}
H^{\mathrm{G}}(t,\tau) &= e^{-\int_\tau^t \left(\frac{\sqrt{\mu}}{2}\tanh\left(\frac{\sqrt{\mu}}{2}(T-s)\right) + \frac{3\sqrt{\mu}}{2}\coth\left(\frac{\sqrt{\mu}}{2}(T-s)\right)\right) ds} \\
&= e^{\left[3\log\left(\sinh\left(\frac{\sqrt{\mu}}{2}(T-s)\right)\right) + \log\left(\cosh\left(\frac{\sqrt{\mu}}{2}(T-s)\right)\right)\right]_\tau^t} \\
&= \frac{\sinh^3\left(\frac{\sqrt{\mu}}{2}(T-t)\right)\cosh\left(\frac{\sqrt{\mu}}{2}(T-t)\right)}{\sinh^3\left(\frac{\sqrt{\mu}}{2}(T-\tau)\right)\cosh\left(\frac{\sqrt{\mu}}{2}(T-\tau)\right)}.
\end{aligned}$$

Note that the differential kernel of the unified AGM ODE is (see Appendix D.2)

$$H^{\mathrm{F}}(t,\tau) = \frac{\sinh^3\left(\frac{\sqrt{\mu}}{2}\tau\right)\cosh\left(\frac{\sqrt{\mu}}{2}\tau\right)}{\sinh^3\left(\frac{\sqrt{\mu}}{2}t\right)\cosh\left(\frac{\sqrt{\mu}}{2}t\right)}.$$

Now, we can observe that there is an anti-transpose relationship (16) between these ODEs.

F.4. Proof of Theorem 6.1

Clearly, the unified AGM-G ODE (17) has a unique solution $X(t)$ in $C^1([0,T),\mathbb{R}^n)$.[24] We can continuously extend this solution to $t = T$ with $\dot{X}(T) = 0$ and $\ddot{X}(T) = \lim_{t\to T^-} \frac{\dot{X}(t)}{t-T} = \frac{1}{2}\nabla f(X(T))$ (see Appendix F.5). Denote this extended solution by $X : [0,T] \to \mathbb{R}^n$.

[24] Sketch of the proof: For any $\epsilon \in (0, T/2)$, the existence and uniqueness of solution on $[0, T-\epsilon]$ follows from Cauchy-Lipschitz theorem (Teschl, 2012, Theorem 25). Paste these solutions on $[0,T) = \cup_{\epsilon\in(0,T/2)}[0,T-\epsilon)$.

For convenience, we assume $\mu > 0$ (the case $\mu = 0$ can be handled easily). Define a continuous-time energy function $\mathcal{E}(t)$ as

$$\begin{aligned}\mathcal{E}(t) = &\frac{4}{(T-t)^2}\operatorname{cschc}^2\left(\frac{\sqrt{\mu}}{2}(T-t)\right)(f(X(t)) - f(X(T)))\\ &- \frac{8}{(T-t)^4}\operatorname{cschc}^4\left(\frac{\sqrt{\mu}}{2}(T-t)\right)\|X(t) - X(T)\|^2\\ &+ \frac{8}{(T-t)^4}\operatorname{cschc}^2\left(\frac{\sqrt{\mu}}{2}(T-t)\right)\operatorname{cothc}^2\left(\frac{\sqrt{\mu}}{2}(T-t)\right)\\ &\times\left\|X(t) + \frac{T-t}{2}\operatorname{tanhc}\left(\frac{\sqrt{\mu}}{2}(T-t)\right)\dot{X}(t) - X(T)\right\|^2.\end{aligned} \tag{112}$$

For simplicity, we denote $X(t)$ by X, and $X(T)$ by x^T. We also omit the input $\frac{\sqrt{\mu}}{2}(T-t)$ of each hyperbolic function. For example, we write the unified AGM-G ODE (17) as

$$\ddot{X} + \left(\frac{\sqrt{\mu}}{2}\tanh + \frac{3\sqrt{\mu}}{2}\coth\right)\dot{X} + \nabla f(X) = 0$$

and the continuous-time energy function (112) as

$$\mathcal{E}(t) = \mu^2\operatorname{csch}^4\left(\frac{\sinh^2}{\mu}\left(f(X) - f\left(x^T\right)\right) - \frac{1}{2}\left\|X - x^T\right\|^2 + \frac{\cosh^2}{2}\left\|X + \frac{\tanh}{\sqrt{\mu}}\dot{X} - x^T\right\|^2\right).$$

Then, we have

$$\begin{aligned}&\frac{\sinh^4}{\mu^2}\dot{\mathcal{E}}(t)\\ &= \sinh^4\frac{d}{dt}\left\{\operatorname{csch}^4\right\}\left(\frac{\sinh^2}{\mu}\left(f(X) - f\left(x^T\right)\right) - \frac{1}{2}\left\|X - x^T\right\|^2 + \frac{\cosh^2}{2}\left\|X + \frac{\tanh}{\sqrt{\mu}}\dot{X} - x^T\right\|^2\right)\\ &\quad + \frac{d}{dt}\left\{\frac{\sinh^2}{\mu}\left(f(X) - f\left(x^T\right)\right) - \frac{1}{2}\left\|X - x^T\right\|^2 + \frac{\cosh^2}{2}\left\|X + \frac{\tanh}{\sqrt{\mu}}\dot{X} - x^T\right\|^2\right\}\\ &= 2\sqrt{\mu}\coth\left(\frac{\sinh^2}{\mu}\left(f(X) - f\left(x^T\right)\right) - \frac{1}{2}\left\|X - x^T\right\|^2 + \frac{\cosh^2}{2}\left\|X + \frac{\tanh}{\sqrt{\mu}}\dot{X} - x^T\right\|^2\right)\\ &\quad - \frac{\sinh\cosh}{\sqrt{\mu}}\left(f(X) - f\left(x^T\right)\right) + \frac{\sinh^2}{\mu}\left\langle\nabla f(X), \dot{X}\right\rangle - \left\langle X - x^T, \dot{X}\right\rangle\\ &\quad - \frac{\sqrt{\mu}\sinh\cosh}{2}\left\|X + \frac{\tanh}{\sqrt{\mu}}\dot{X} - x^T\right\|^2 + \cosh^2\left\langle X + \frac{\tanh}{\sqrt{\mu}}\dot{X} - x^T, -\dot{X} - \frac{\tanh}{\sqrt{\mu}}\nabla f(X)\right\rangle,\end{aligned}$$

where we used

$$\begin{aligned}\frac{d}{dt}\left\{X + \frac{\tanh}{\sqrt{\mu}}\dot{X} - x^T\right\} &= \frac{\tanh}{\sqrt{\mu}}\ddot{X} + \left(1 - \frac{1}{2}\operatorname{sech}^2\right)\dot{X}\\ &= \left(-\frac{1}{2}\tanh^2 - \frac{1}{2} - \frac{1}{2}\operatorname{sech}^2\right)\dot{X} - \frac{\tanh}{\sqrt{\mu}}\nabla f(X)\\ &= -\dot{X} - \frac{\tanh}{\sqrt{\mu}}\nabla f(X)\end{aligned}$$

for the last equality. We further simplify as

$$\begin{aligned}\frac{\sinh^4}{\mu^2}\dot{\mathcal{E}}(t) = &\frac{2\sinh\cosh}{\sqrt{\mu}}\left(f(X) - f\left(x^T\right)\right) - \sqrt{\mu}\coth\left\|X - x^T\right\|^2\\ &+ \sqrt{\mu}\coth\cosh^2\left(\left\|X - x^T\right\|^2 + \frac{\tanh^2}{\mu}\left\|\dot{X}\right\|^2 + \frac{2\tanh}{\sqrt{\mu}}\left\langle X - x^T, \dot{X}\right\rangle\right)\end{aligned}$$

$$
\begin{aligned}
&- \frac{\sinh\cosh}{\sqrt{\mu}}\left(f(X)-f\left(x^T\right)\right)+\frac{\sinh^2}{\mu}\left\langle\nabla f(X),\dot{X}\right\rangle-\left\langle X-x^T,\dot{X}\right\rangle\\
&- \frac{\sqrt{\mu}\sinh\cosh}{2}\left(\left\|X-x^T\right\|^2+\frac{\tanh^2}{\mu}\left\|\dot{X}\right\|^2+\frac{2\tanh}{\sqrt{\mu}}\left\langle X-x^T,\dot{X}\right\rangle\right)\\
&- \cosh^2\left(\left\langle X-x^T,\dot{X}\right\rangle+\frac{\tanh}{\sqrt{\mu}}\left\|\dot{X}\right\|^2\right.\\
&\qquad\left.+\frac{\tanh}{\sqrt{\mu}}\left\langle X-x^T,\nabla f(X)\right\rangle+\frac{\tanh^2}{\mu}\left\langle\dot{X},\nabla f(X)\right\rangle\right)\\
=&\left(\frac{2\sinh\cosh}{\sqrt{\mu}}-\frac{\sinh\cosh}{\sqrt{\mu}}\right)\left(f(X)-f\left(x^T\right)\right)\\
&+\left(-\sqrt{\mu}\coth+\sqrt{\mu}\coth\cosh^2-\frac{\sqrt{\mu}\sinh\cosh}{2}\right)\left\|X-x^T\right\|^2\\
&+\left(\frac{\sinh\cosh}{\sqrt{\mu}}-\frac{\sinh^2\tanh}{2\sqrt{\mu}}-\frac{\sinh\cosh}{\sqrt{\mu}}\right)\left\|\dot{X}\right\|^2\\
&+\left(2\cosh^2-1-\sinh^2-\cosh^2\right)\left\langle X-x^T,\dot{X}\right\rangle\\
&+\left(\frac{\sinh^2}{\mu}-\frac{\sinh^2}{\mu}\right)\left\langle\nabla f(X),\dot{X}\right\rangle\\
&-\frac{\sinh\cosh}{\sqrt{\mu}}\left\langle X-x^T,\nabla f(X)\right\rangle\\
=&\frac{\sinh\cosh}{\sqrt{\mu}}\left(f(X)-f\left(x^T\right)\right)\\
&+\frac{\sqrt{\mu}\sinh\cosh}{2}\left\|X-x^T\right\|^2-\frac{\sinh^2\tanh}{2\sqrt{\mu}}\left\|\dot{X}\right\|^2-\frac{\sinh\cosh}{\sqrt{\mu}}\left\langle X-x^T,\nabla f(X)\right\rangle.
\end{aligned}
$$

It follows from the μ-strong convexity of f that $f(X)-f\left(x^T\right)\le\left\langle X-x^T,\nabla f(X)\right\rangle-\frac{\mu}{2}\left\|X-x^T\right\|^2$. Thus, we have

$$
\begin{aligned}
\frac{\sinh^4}{\mu^2}\dot{\mathcal{E}}(t)\le&\frac{\sinh\cosh}{\sqrt{\mu}}\left(\left\langle X-x^T,\nabla f(X)\right\rangle-\frac{\mu}{2}\left\|X-x^T\right\|^2\right)\\
&+\frac{\sqrt{\mu}\sinh\cosh}{2}\left\|X-x^T\right\|^2-\frac{\sinh^2\tanh}{2\sqrt{\mu}}\left\|\dot{X}\right\|^2-\frac{\sinh\cosh}{\sqrt{\mu}}\left\langle X-x^T,\nabla f(X)\right\rangle\\
=&-\frac{\sinh^2\tanh}{2\sqrt{\mu}}\left\|\dot{X}\right\|^2\\
\le&\,0.
\end{aligned}
$$

Therefore, the energy function $\mathcal{E}(t)$ is non-increasing. By L'Hôpital's rule, we have

$$
\begin{aligned}
\lim_{t\to T^-}\frac{f(X(t))-f(X(T))}{(T-t)^2}&=\lim_{t\to T^-}\frac{1}{2}\left\langle\frac{\dot{X}(t)}{t-T},\nabla f(X)\right\rangle=\frac{1}{4}\left\|\nabla f(X(T))\right\|^2\\
\lim_{t\to T^-}\frac{X(t)-X(T)}{(T-t)^2}&=\lim_{t\to T^-}\frac{\dot{X}(t)}{2(t-T)}=\frac{1}{4}\nabla f(X(T)).
\end{aligned}
$$

It follows from $\mathrm{cschc}(0)=\mathrm{cothc}(0)=1$ that

$$
\begin{aligned}
&\lim_{t\to T^-}\mathcal{E}(t)\\
=&\lim_{t\to T^-}\left(4\cdot\frac{f(X(t))-f(X(T))}{(T-t)^2}-8\left\|\frac{X(t)-X(T)}{(T-t)^2}\right\|^2+8\left\|\frac{X(t)-X(T)}{(T-t)^2}-\frac{\dot{X}(t)}{2(t-T)}\right\|^2\right)
\end{aligned}
$$

$$= \|\nabla f(X(T))\|^2 - \frac{1}{2}\|\nabla f(X(T))\|^2 + 0$$
$$= \frac{1}{2}\|\nabla f(X(T))\|^2.$$

Writing $\lim_{t\to T^-}\mathcal{E}(t) \le \mathcal{E}(0)$ explicitly, we obtain

$$\|\nabla f(X(T))\|^2 \le \frac{8}{T^2}\operatorname{cschc}^2\left(\frac{\sqrt{\mu}}{2}T\right)\left(f(x_0) - f(X(T)) + \frac{\mu}{2}\|x_0 - X(T)\|^2\right)$$
$$\le \frac{8}{T^2}\operatorname{cschc}^2\left(\frac{\sqrt{\mu}}{2}T\right)\sup_x\left\{f(x_0) - f(x) + \frac{\mu}{2}\|x_0 - x\|^2\right\}.$$

Since cschc^2 is decreasing on $[0,\infty)$, this implies that the unified AGM-G ODE (69) reduces the squared gradient norm with an $O\left(1/T^2\right)$ convergence rate regardless of the value of $\mu \ge 0$. When $\mu > 0$, since $\frac{1}{T^2}\operatorname{cschc}^2\left(\frac{\sqrt{\mu}}{2}T\right) \sim \mu e^{-\sqrt{\mu}T}$ as $T\to\infty$, the unified AGM-G ODE reduces the squared gradient norm with an $O\left(e^{-\sqrt{\mu}T}\right)$ convergence rate. Combining these bounds, we conclude that the unified AGM-G ODE reduces the squared gradient norm with the following convergence rate:

$$\|\nabla f(X(T))\|^2 \le O\left(\min\left\{1/T^2, e^{-\sqrt{\mu}T}\right\}\right).$$

This completes the proof.

Remark F.1. One might expect that the anti-transposed dynamics of AGM-SC ODE reduce the gradient norm with an $O(e^{-\sqrt{\mu}t})$ convergence rate. However, the argument in this subsection cannot be seamlessly applied to this dynamics. The differential kernel of AGM-SC ODE is $H^{\mathrm{SC}}(t,\tau) = e^{2\sqrt{\mu}(\tau-t)}$ (see Appendix B.2.3). Because $H^{\mathrm{SC}}(t,\tau)$ is anti-symmetric, the anti-transposed dynamics of AGM-SC ODE is itself:

$$\ddot{X} + 2\sqrt{\mu}\dot{X} + \nabla f(X) = 0, \tag{113}$$

In the proof of Theorem 6.1, the property $\dot{X}(T) = 0$ is essentially used. However, the solution to (113) does not satisfy this property.

F.5. Computing $\dot{X}(T)$ and $\ddot{X}(T)$

For simplicity, we assume that the limits $\lim_{t\to T^-}\dot{X}(T)$ and $\lim_{t\to T^-}\ddot{X}(T)$ exist.[25] Consider the energy function

$$\mathcal{E}(t) = \frac{1}{2}\left\|\dot{X}(t)\right\|^2 + (f(X(t)) - f(x^*)) + \int_0^t\left[\frac{\sqrt{\mu}}{2}\tanh\left(\frac{\sqrt{\mu}}{2}(T-s)\right) + \frac{3}{T-s}\operatorname{cothc}\left(\frac{\sqrt{\mu}}{2}(T-s)\right)\right]\left\|\dot{X}(s)\right\|^2 ds. \tag{114}$$

Then, it is easy to show that $\mathcal{E}(t) = \mathcal{E}(0)$ for all $t \in [0,T)$. Because the terms $\frac{1}{2}\left\|\dot{X}(t)\right\|^2$ and $f(X(t)) - f(x^*)$ are non-negtive, we have

$$\int_0^T\left[\frac{\sqrt{\mu}}{2}\tanh\left(\frac{\sqrt{\mu}}{2}(T-s)\right) + \frac{3}{T-s}\operatorname{cothc}\left(\frac{\sqrt{\mu}}{2}(T-s)\right)\right]\left\|\dot{X}(s)\right\|^2 ds < \infty.$$

This implies $\lim_{t\to T^-}\dot{X}(t) = 0$. By L'Hôpital's rule, we obtain that

$$\lim_{t\to T^-}\left[\frac{\sqrt{\mu}}{2}\tanh\left(\frac{\sqrt{\mu}}{2}(T-t)\right) + \frac{3}{T-t}\operatorname{cothc}\left(\frac{\sqrt{\mu}}{2}(T-t)\right)\right]\dot{X}(t) = -3\ddot{X}(T).$$

Now, we have

$$0 = \lim_{t\to T^-}\left\{\ddot{X}(t) + \left[\frac{\sqrt{\mu}}{2}\tanh\left(\frac{\sqrt{\mu}}{2}(T-t)\right) + \frac{3}{T-t}\operatorname{cothc}\left(\frac{\sqrt{\mu}}{2}(T-t)\right)\right]\dot{X}(t) + \nabla f(X(t))\right\}$$
$$= -2\ddot{X}(T) + \nabla f(X(T)).$$

Thus, we have $\ddot{X}(T) = \frac{1}{2}\nabla f(X(T))$.

[25] The proof to prove the existence of these limits is similar to that in (Suh et al., 2022, Appendix D.3), so we omit it.