MATHEMATICAL ENGINEERING
TECHNICAL REPORTS

Deriving Optimal Rates of Continuous-time Accelerated First-order Methods via Performance Estimation Problems

Kansei USHIYAMA, Shun SATO, Takayasu MATSUO

(Communicated by Takayasu MATSUO)

METR 2024–02 February 2024

DEPARTMENT OF MATHEMATICAL INFORMATICS
GRADUATE SCHOOL OF INFORMATION SCIENCE AND TECHNOLOGY
THE UNIVERSITY OF TOKYO
BUNKYO-KU, TOKYO 113-8656, JAPAN

WWW page: https://www.keisu.t.u-tokyo.ac.jp/research/techrep/

Deriving Optimal Rates of Continuous-time Accelerated First-order Methods via Performance Estimation Problems

Kansei USHIYAMA, Shun SATO, Takayasu MATSUO*

February 2024

Abstract

In this study, optimal or near-optimal convergence rates are proved for continuous-time models of accelerated first-order methods. This encompasses not only a multitude of established results but also reveals some new rates. Specifically, for the ordinary differential equation (ODE) model representing the information-theoretic exact method by Taylor–Drori (2022), which is recognized as the fastest and most efficient lower bound-achieving method, these rates have been newly and rigorously established. The basis of this derivation lies in the continuous adaptation of the performance estimation problem, initially introduced by Drori and Teboulle (2014). This study completes the overview of ODE rates.

1 Introduction

We consider an unconstrained convex optimization problem:

$$\underset{x\in\mathbb{R}^d}{\text{minimize}}\, f(x).$$

For this problem, various first-order methods that span from the classical gradient descent to several advanced accelerated methods (originating in (Nesterov, 1983)) have been devised. Their convergence has been intensively examined through various performance measures such as $f(x^{(k)}) - f^\star$, $\|\nabla f(x^{(k)})\|$, and $\|x^{(k)} - x^\star\|$.

Convergence analyses are usually complex. However, the recent shift to a more intuitive approach using ordinary differential equations (ODEs) instead of discrete optimization methods has gained significant traction. This shift offers a more accessible understanding and theoretical insights, drawing from physics and dynamic systems theory. This approach has been known for a long time but has gained significant attention after the seminal work (Su et al., 2014), where an ODE was conceptualized as a continuous-time limit of Nesterov's accelerated gradient descent. An important tool in convergence analysis is the Lyapunov function, which establishes the convergence rate of the ODE. This approach has been validated by numerous subsequent studies (referenced in Section 1.3.1). One limitation of this approach is that Lyapunov functions are typically identified through trial and error. Recent research has focused on the automatic construction of Lyapunov functions (Suh et al., 2022; Moucer et al., 2023), yet these methods do not encompass all optimization ODEs, nor do they consistently achieve the optimal convergence rates expected from discrete method lower bounds.

A recent innovation is the systematic approach for assessing the worst-case performance of discrete-time first-order methods, extensively studied as performance estimation problems (PEP) (Drori & Teboulle, 2014). This approach formalizes the worst-case error of an optimization method as a solution to another optimization problem. By resolving this problem, occasionally with computer assistance, we can determine the desired worst-case performance or its upper bound. Its continuous-time counterpart, termed the "continuous PEP," was later proposed in (Kim & Yang, 2023b). This innovation reveals the convergence rates of optimization ODEs expressible in integro-differential equations (IDEs). This development is groundbreaking as it introduces a second systematic way for analyzing ODEs, following the Lyapunov approach. It is anticipated to lead to numerous new findings within this framework. However, the continuous PEP has its limitations; it is applicable to a restricted range of ODEs, and sometimes the convergence rates it yields are not optimal.

In this paper, we propose an advanced continuous PEP framework designed to overcome these limitations. To distinguish it from the original continuous PEP (Kim & Yang, 2023b) referred as "cPEP," we label our new version as "cPEP⋆".

*Department of Mathematical Informatics, Graduate School of Information Science and Technology, The University of Tokyo (ushiyama-kansei074@g.ecc.u-tokyo.ac.jp).

1.1 Contributions

New flexible framework cPEP*. This is a new, more versatile continuous PEP. It is adjustable for handling not only $f - f^\star$ and $\|\nabla f\|$ as in (Kim & Yang, 2023b) but also $\|x - x^\star\|$, and for optimizing the overall discussion to obtain optimal rates. Furthermore, it covers ODEs that cPEP does not (ODEs with shifted gradients.)

Discovery of new optimal (or near-optimal) rates. The newly identified rates pertain to ODEs for strongly convex functions, which are expected to be the fastest to date (the ODEs for TMM and ITEM; see Section 1.3.1). Also new stronger rates are obtained for some high-resolution ODEs. cPEP* also recovers other known optimal rates successfully.

New Lyapunov functions. As an unexpected secondary product, new Lyapunov functions are found via cPEP*.

This research completes the overview of ODE rates (as detailed in Table 2). Including the established rates, all rates are optimal or nearly so, aligning closely with the optimal rates achievable in discrete methods.

1.2 Notations and settings

In our notation, $\langle \cdot, \cdot \rangle$ and $\|\cdot\|$ denote the inner product and norm in Euclidean space $\mathbb{R}^d$, respectively. $\mathcal{F}_{\mu,L}$ represents the set of μ-strongly convex and L-smooth functions, allowing for cases where $\mu = 0$ or $L = \infty$ (i.e., f is just differentiable). For f, we assume the existence of a minimal solution $x^\star$ and its corresponding minimum value $f^\star$.

1.3 Related works

1.3.1 ODE approach and analysis

The concept of modeling momentum methods as ODEs dates back to the studies of (Polyak, 1964), with an important milestone being the work of (Su et al., 2014). In this work, they derived a second-order ODE model of Nesterov's accelerated gradient method for convex functions (AGM) (Nesterov, 1983), later extended for mirror descent (Krichene et al., 2015). Numerous variants of the AGM ODE have been studied, including ODEs with different damping mechanisms (Attouch & Cabot, 2017; May, 2017; Aujol & Dossal, 2017; Attouch et al., 2019), Hessian damping (Attouch et al., 2016; 2022c;b), perturbations, and Tikhonov regularization (Attouch et al., 2018a;b; Boţ et al., 2021; Attouch et al., 2022a).

A systematic way for modeling optimization methods is the introduction of "high-resolution ODEs." These ODEs incorporate step-size information, enabling a more precise reproduction of the outputs of the target optimization method (Shi et al., 2022; Lu, 2022; Sun et al., 2020; Chen et al., 2023). This enables us to distinguish Nesterov's accelerated gradient method for strongly convex functions (AGM-SC) (Nesterov, 2018) and Polyak's heavy-ball method (Polyak, 1964) in continuous-time modeling; the same ODE is obtained for each method via low-resolution ODEs (cf. (Wilson et al., 2021)). This technique aligns with the concept of the modified equation in numerical analyses (Sanz Serna & Zygalakis, 2021). They are particularly effective for analyzing the convergence of the gradient norm (Chen et al., 2022; Maskan et al., 2023).

Other continuous-time models use approaches such as the duality gap (Diakonikolas & Orecchia, 2019) and the Bregman Lagrangian (Wibisono et al., 2016; Wilson et al., 2021).

The standard practice for analyzing ODE models involves using Lyapunov functions, which explicitly state the expected convergence rates. This method was initially applied to second-order ODE models (Su et al., 2014) and later refined (Wilson et al., 2021). For discrete-time algorithms, a similar analytical framework can be applied, as reviewed in (Bansal & Gupta, 2019). The main advantage of Lyapunov functions is their simplicity; once identified, they almost complete the proof of the rate. However, finding suitable Lyapunov functions has largely depended on the expertise of researchers in each study, particularly for complicated cases such as high-resolution ODEs. While there are a few studies on the automatic construction of Lyapunov functions (Suh et al., 2022; Moucer et al., 2023), their scope remains limited, and there is no guarantee of finding an optimal Lyapunov function that states the desired rate. For discrete-time algorithms, computer-assisted methods for finding Lyapunov functions have been investigated (Taylor et al., 2018; Taylor & Bach, 2019).

1.3.2 Performance estimation problem

PEP was originally proposed in (Drori & Teboulle, 2014). PEP is a problem in function space, defined for a fixed integer K.

$$\begin{aligned} \underset{f, x^{(0)}, \ldots, x^{(K)}}{\text{maximize}} \quad & f(x^{(K)}) - f^\star \\ \text{subject to} \quad & f \text{ is convex and } L\text{-smooth}, \end{aligned}$$

$$x^{(k)} \text{ is generated by a first-order method,}$$
$$\text{conditions on initial data.}$$

It is difficult to solve due to f moving in an infinite-dimensional space. However, PEP can be rewritten as a finite-dimensional semidefinite program (SDP) using convex interpolation theory (Taylor et al., 2017a). It can then be solved by computer-based solutions (libraries are found in (Taylor et al., 2017b; Goujaud et al., 2022)). The numerical results from these solutions provide valuable insights for constructing fast first-order methods. Several optimal methods have been derived through PEP, such as optimized gradient methods (OGM) (Kim & Fessler, 2016), OGM-G (Kim & Fessler, 2021), and the information-theoretic exact method (ITEM) (Taylor & Drori, 2022), along with convergence analyses performed by analytically processing the PEP.

Inspired by the original work (Drori & Teboulle, 2014), another branch of the PEP is the integral quadratic constraint (IQC) technique. This framework employs control theory perspectives to analyze optimization methods, particularly focusing on Lyapunov functions (Lessard et al., 2016; Fazlyab et al., 2018). Using the IQC method, the triple momentum method (TMM) was designed (Van Scoy et al., 2018). TMM achieves an optimal convergence rate for strongly convex functions, up to a constant factor.

1.4 Brief review on cPEP

In (Kim & Yang, 2023b), the continuous-time models expressed as IDEs were considered:

$$\dot{x}(t) = -\int_0^t H(t,\tau)\nabla f(x(\tau))\mathrm{d}\tau,$$

where $H(\cdot,\cdot)$ is a kernel satisfying certain conditions. By selecting different kernels, the equation can represent various ODEs. For this equation, a continuous PEP framework was established, focusing on terms such as $f(x(T)) - f^\star$ and $\|\nabla f(x(T)\|$, and applied to analyze the convergence of a variety of ODE models by rewriting them into the IDE form. This includes ODEs for AGM, AGM-SC, TMM, ITEM/ITEM-G[1], OGM-G, and unified AGM/AGM-G (Kim & Yang, 2023a). The critical concept in these analyses is viewing the continuous problem as an infinite-dimensional SDP, enabling a discussion similar to that of the discrete PEP.

Despite the significant contributions of this pioneering work, there are some limitations. First, the IDE may not accommodate methods involving shifted gradients such as $\nabla f(x(t) + \beta(t)\dot{x}(t))$, which are important in the high-resolution ODE context, and sharpness-aware minimizations (Foret et al., 2021) which have gained attention in recent years for their good generalization of neural networks. Second, directly handling strongly convex conditions using IDE is challenging. As a result, the cPEP framework primarily addresses convex functions, with strongly convex functions being reparametrized into convex functions. Consequently, the rate bounds obtained are in terms of $f(x(T)) - f^\star - \frac{\mu}{2}\|x(T) - x^\star\|^2$ rather than $f(x(T)) - f^\star$.

1.5 Our setting for new continuous PEP

In this study, we adopted a different strategy. We start with a specific ODE:

$$\ddot{x}(t) + \alpha(t)\dot{x}(t) + 2\nabla f(x(t) + \beta(t)\dot{x}(t)) = 0, \tag{1}$$

where $\alpha, \beta : [0,T] \to \mathbb{R}$ are given functions. This choice successfully includes all concrete ODEs considered in (Kim & Yang, 2023b) and various ODEs with shifted gradients. This approach is notable as it considers the effect of the Hessian of f, which is frequently considered in the high-resolution ODEs. To demonstrate this phenomenon, let us consider cases where $\alpha(t)$ and $\beta(t)$ are constants. Then, by defining $y(t) \coloneqq x(t) + \beta\dot{x}(t)$, we can rewrite the ODE as follows:

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

This corresponds to the high-resolution ODE for AGM-SC proposed in (Shi et al., 2022). In Appendix E, we consider the high-resolution ODE of the heavy-ball method and AGM-SC, which is written in the above form. Proven rates are faster than that shown in (Shi et al., 2022) using Lyapunov functions.

[1]Derived in the Appendix of (Kim & Yang, 2023b) as a type of dual form of the ITEM ODE. This model arises in discussions of continuous-time models and currently lacks a corresponding discrete method.

The inclusion of the coefficient 2 on ∇f in (1) is a tactical decision to standardize the timescale. In continuous models, convergence rates can be arbitrarily accelerated by rescaling time. However, during discretization, this is countered by numerical instability, rendering such rescaling largely irrelevant to the convergence rate of the discrete method (Ushiyama et al., 2022). The coefficient 2 was selected to align the scale with that of known optimal methods such as OGM/OGM-G, TMM, and ITEM/ITEM-G. The AGM ODE (2), AGM-SC ODE (4), and unified AGM ODE (6) should be accelerated by a factor of $\sqrt{2}$ to match their coefficient to 2.

Our cPEP* is demonstrated in Section 2 using the ODE model of TMM as a case study. This gives a new, desired optimal rate of the ODE. In Section 3, we outline the main theorems, including the case of TMM. The results suggest the potential for constructing novel optimal methods through strategic discretization of optimal ODEs. As a secondary outcome, new Lyapunov functions for the TMM and ITEM ODEs are also identified. This contributes to a comprehensive understanding of ODE model analysis from both PEP and Lyapunov perspectives.

1.6 Key ODE models and summary of results

For reader reference, various ODE models encompassed by our PEP* are listed:

- AGM ODE (Su et al., 2014)
$$\ddot{x} + \frac{3}{t}\dot{x} + \nabla f(x) = 0, \tag{2}$$
- OGM ODE (obtained in the limit of $\mu \to 0$ in (7))
$$\ddot{x} + \frac{3}{t}\dot{x} + 2\nabla f(x) = 0, \tag{3}$$
- AGM-SC ODE (cf. (Wilson et al., 2021))
$$\ddot{x} + 2\sqrt{\mu}\dot{x} + \nabla f(x) = 0, \tag{4}$$
- TMM ODE (Kim & Yang, 2023b)
$$\ddot{x} + 3\sqrt{\mu}\dot{x} + 2\nabla f(x) = 0, \tag{5}$$
- Unified AGM ODE (Kim & Yang, 2023a)
$$\ddot{x} + \frac{\sqrt{\mu}}{2}\left(\tanh\left(\frac{\sqrt{\mu}}{2}t\right) + 3\coth\left(\frac{\sqrt{\mu}}{2}t\right)\right)\dot{x} + \nabla f(x) = 0, \tag{6}$$
- ITEM ODE (Kim & Yang, 2023b)
$$\ddot{x} + 3\sqrt{\mu}\coth(\sqrt{\mu}t)\dot{x} + 2\nabla f(x) = 0, \tag{7}$$
- OGM-G ODE (Suh et al., 2022)
$$\ddot{x} + \frac{3}{S-t}\dot{x} + 2\nabla f(x) = 0, \tag{8}$$
- ITEM-G ODE (Kim & Yang, 2023b)
$$\ddot{x} + 3\sqrt{\mu}\coth(\sqrt{\mu}(S-t))\dot{x} + 2\nabla f(x) = 0. \tag{9}$$

Tables 1–2 compile the rates and their corresponding methods or ODEs. In discrete methods, exact lower bounds have been derived in (Drori, 2017; Drori & Taylor, 2022). For convex f in terms of $f - f^\star$, OGM precisely matches the worst-case rate with the lower bound. In strongly convex f, ITEM achieves the lower bound for $\|x - x^\star\|$.

In Table 2, rates labeled as **new** are new findings. We will show the rates and their achieving ODEs below, via cPEP*. cPEP* reaffirms known results, ensuring consistency with related studies. Discussing lower bounds for ODE model rates is challenging due to potential time rescalings, and no definitive results are known. However, the table highlights rates corresponding to optimal discrete methods.

Table 1: Convergence rates of fastest first-order method for L-smooth f. All of them are optimal up to constant, except for † exactly giving lower bound. Achieving methods are $\langle 1\rangle$ OGM (Kim & Fessler, 2016), $\langle 2\rangle$ combination of OGM-G and AGM (Kim & Fessler, 2021; Nesterov et al., 2021), $\langle 4\rangle$ ITEM (Taylor & Drori, 2022); $\langle 3\rangle$ and $\langle 5\rangle$ come from $\langle 4\rangle$ using the μ-strongly convexity and L-smoothness.

	convex	μ-strongly convex
$f(x^{(k)}) - f^\star$	$\mathrm{O}\big(1/k^2\big)^\dagger$ $\langle 1\rangle$	$\mathrm{O}\Big(\big(1-\sqrt{\frac{\mu}{L}}\big)^{2k}\Big)$ $\langle 3\rangle$
$\|x^{(k)} - x^\star\|$	—	$\mathrm{O}\Big(\big(1-\sqrt{\frac{\mu}{L}}\big)^{k}\Big)^\dagger$ $\langle 4\rangle$
$\|\nabla f(x^{(k)})\|$	$\mathrm{O}\big(1/k^2\big)$ $\langle 2\rangle$	$\mathrm{O}\Big(\big(1-\sqrt{\frac{\mu}{L}}\big)^{k}\Big)$ $\langle 5\rangle$

Table 2: Convergence rates of ODE (1) for ∞-smooth f. Achieving concrete ODEs are $\langle 1\rangle$ OGM ODE, $\langle 2\rangle$ combination of OGM ODE and OGM-G ODE, $\langle 3\rangle$ ODE in Theorem 3.3, $\langle 4\rangle$ ITEM ODE, and $\langle 5\rangle$ ITEM-G ODE. The symbol $\star$ means the rate corresponds to the lower bound of discrete first-order methods, "**new**" means they are new in this study, and "(L)" indicates the additional assumption $L < \infty$. ε is an arbitrarily small constant (see Theorem 3.1).

	convex	μ-strongly convex
$f(x(t)) - f^\star$	$\mathrm{O}\big(1/t^2\big)^\star$ $\langle 1\rangle$	$\mathrm{O}\big(\mathrm{e}^{-(2\sqrt{\mu}-\varepsilon)t}\big)^{\mathbf{new}}$ $\langle 3\rangle$ $\mathrm{O}\big(\mathrm{e}^{-2\sqrt{\mu}t}\big)^{\star,\mathbf{new}}$ $\langle 4\rangle(L)$
$\|x(t) - x^\star\|$	—	$\mathrm{O}\big(\mathrm{e}^{-\sqrt{\mu}t}\big)^{\star,\mathbf{new}}$ $\langle 4\rangle$
$\|\nabla f(x(t))\|$	$\mathrm{O}\big(1/t^2\big)^\star$ $\langle 2\rangle$	$\mathrm{O}\big(\mathrm{e}^{-\sqrt{\mu}t}\big)^\star$ $\langle 5\rangle$

The situation regarding the rates for $f - f^\star$ of strongly convex f is complicated. In the discrete case, the (up-to-constant) optimal rate $\mathrm{O}\Big((1-\sqrt{\mu/L})^{2k}\Big)$ is obtained using ITEM (the optimal method for $\|x - x^\star\|$) and L-smoothness $f(x) - f^\star \leq \frac{L}{2}\|x - x^\star\|^2$. This leaves L in the constant factor, which is not optimal. The existence of a better method than ITEM is suggested in (Taylor & Drori, 2022), but this method has not yet been discovered. In the continuous case, L-smoothness is not typically necessary in the analysis. Moreover, we reveal that ITEM ODE has a corresponding rate to the discrete ITEM with respect to $\|x - x^\star\|$ without L-smoothness. However, for the rates on $f - f^\star$, the direct cPEP* result is slightly weak, including an arbitrarily small constant ε (the first rate in the table). If we use the same strategy as in the discrete case, the corresponding (up-to-constant) optimal rate $\mathrm{O}\big(\mathrm{e}^{-2\sqrt{\mu}t}\big)$ is obtained (i.e., the second rate). Although this might seem stronger, it is not necessarily so because it includes the unnecessary L in the constant factor in the order. The former rate might suggest the possibility of new method that does not have L in the constant factor, which is a discretization of the ITEM ODE with a slightly weakened damping term.

2 Continuous performance estimation problem

To demonstrate our continuous PEP formulation, we consider an example:

$$
\begin{aligned}
\underset{\substack{f\in\mathcal{F}_{\mu,L},\\ X\in C^2([0,T])}}{\text{maximize}} \quad & \mu\left\|x(T) + \frac{1}{2\sqrt{\mu}}\dot{x}(T) - x^\star\right\|^2 \\
\text{subject to} \quad & \ddot{x}(t) + 3\sqrt{\mu}\dot{x}(t) + 2\nabla f(x(t)) = 0, && (10)\\
& f(x(0)) - f^\star \leq R_1, && (11)\\
& \|x(0) - x^\star\|^2 \leq R_2, \quad \dot{x}(0) = 0, && (12)
\end{aligned}
$$

where T, R_1 and R_2 are prescribed constants.

The eq. (10) models the TMM (5). Solving this maximization problem determines the worst-case ODE performance in terms of the performance measure $\|x(T) + \frac{1}{2\sqrt{\mu}}\dot{x}(T) - x^\star\|^2$. This choice is made to obtain a better rate than that of $\|x(T) - x^\star\|$. This will be elucidated in the subsequent analysis. It interestingly suggests that $x(T) + \frac{1}{2\sqrt{\mu}}\dot{x}(T)$ may be

a more preferable output than $x(T)$ for the continuous-time algorithm (10). This technique has already been utilized in traditional Lyapunov-type analyses. For example, in analyzing AGM-SC ODE (4), the Lyapunov function is defined as $f(x) - f^\star + \frac{\mu}{2}\|x + \frac{1}{\sqrt{2\mu}}\dot{x} - x^\star\|^2$ (Wilson et al., 2021), which ensures the convergence of $x + \dot{x}/\sqrt{2\mu}$ to $x^\star$.

The original problem is too complex to solve directly in $\mathcal{F}_{\mu,L}$, so the aim is to provide an upper bound using the sequence of inequalities: (Original problem) $\leq$ (relaxed problem) $\leq$ (dual problem) $\leq$ (modified minimization problem) $\leq$ (feasible solution). This relaxation and dual strategy is a common approach in PEP, cPEP, and cPEP*; however, the methods of relaxation and duality differ. The first step of this process is detailed in Section 2.1. The second and third in Section 2.2. Then the last in Section 2.3. We only show the outline, where the detail is left to Appendix B.

2.1 Relaxing the problem

To simplify the presentation, the problem is reformulated using $\xi(t) \coloneqq x(t) - x^\star$ before relaxation as

$$\begin{aligned}
\underset{\substack{f\in\mathcal{F}_{\mu,L},\\ \xi,\phi,\gamma}}{\text{maximize}} \quad & \mu\left\|\xi(T) + \frac{1}{2\sqrt{\mu}}\dot{\xi}(T)\right\|^2 \\
\text{subject to} \quad & \phi(t) = f(\xi(t) + x^\star) - f^\star, && (13)\\
& \gamma(t) = \nabla f(\xi(t) + x^\star), && (14)\\
& \ddot{\xi}(t) + 3\sqrt{\mu}\dot{\xi}(t) + 2\gamma(t) = 0, \\
& \phi(0) \leq R_1, \\
& \|\xi(0)\|^2 \leq R_2, \quad \dot{\xi}(0) = 0,
\end{aligned}$$

where $\xi \in C^2([0,T];\mathbb{R}^d)$, $\phi \in C^1([0,T];\mathbb{R})$, $\gamma \in C([0,T];\mathbb{R}^d)$.

The constraint (13) is relaxed to its differentiation $\dot{\phi}(t) = \langle \gamma(t), \dot{\xi}(t)\rangle$. Further relaxations involve the constraints $f \in \mathcal{F}_{\mu,L}$ and (14). It require careful consideration for sharp results. According to convex interpolation theory (Taylor et al., 2017a), these constraints can be rewritten as an inequality without changing the equivalence of the problem.

$$\phi' - \phi'' - \langle \gamma'', \xi' - \xi''\rangle \geq \frac{1}{2(1-\mu/L)}\left(\frac{1}{L}\|\gamma' - \gamma''\|^2 + \mu\|\xi' - \xi''\|^2 - 2\frac{\mu}{L}\langle \gamma'' - \gamma', \xi'' - \xi'\rangle\right) \tag{15}$$

for any $(\phi', \gamma', \xi'), (\phi'', \gamma'', \xi'') \in \mathcal{G}(T) \cup \{(0,0,0)\}$, where $\mathcal{G}(T) \coloneqq \{\, (\phi(t), \gamma(t), \xi(t)) \mid t \in [0,T] \,\}$.

The Lagrange dual problem is considered next. Due to the complexity of introducing Lagrange multipliers over the domain $(\mathcal{G}(T) \cup \{(0,0,0)\})^2$, which is laborious, some constraints are dropped. This is also performed in discrete PEP. The choice of the inequalities to drop strongly affects the sharpness of the result. In this case, as a conclusion, we retain the constraints (15) for $(\phi', \gamma', \xi') = (0,0,0)$, $(\phi'', \gamma'', \xi'') \in \mathcal{G}(T)$ and $(\phi', \gamma', \xi') = (\phi(T), \gamma(T), \xi(T))$, $(\phi'', \gamma'', \xi'') = (0,0,0)$. In this section, the constraints with $L = \infty$ is considered to obtain a better result even if we know $L < \infty$. We discuss what would happen if we utilize the information $L < \infty$ in Section 4 and Appendix B.2.6.

Finally, the relaxed problem is obtained by erasing $\gamma(t)$ and incorporating the ODE (10) into the constraints.

$$\begin{aligned}
\underset{\xi,\phi}{\text{maximize}} \quad & \mu\left\|\xi(T) + \frac{1}{2\sqrt{\mu}}\dot{\xi}(T)\right\|^2 \\
\text{subject to} \quad & \dot{\phi}(t) + \frac{1}{2}\left\langle \ddot{\xi}(t) + 3\sqrt{\mu}\dot{\xi}(t), \dot{\xi}(t)\right\rangle = 0, && (16)\\
& \phi(t) + \frac{1}{2}\left\langle \ddot{\xi}(t) + 3\sqrt{\mu}\dot{\xi}(t), \xi(t)\right\rangle + \frac{\mu}{2}\|\xi(t)\|^2 \leq 0, && (17)\\
& -\phi(T) + \frac{\mu}{2}\|\xi(T)\|^2 \leq 0, && (18)\\
& \phi(0) \leq R_1, \quad \|\xi(0)\|^2 \leq R_2, \quad \dot{\xi}(0) = 0.
\end{aligned}$$

2.2 Dual problem and its modification

Now, let us consider the dual problem. We define the Lagrange function as

$$\begin{aligned}\mathcal{L}(\xi,\phi;\lambda_1,\lambda_2,\lambda_3,\eta_1,\eta_2) \coloneqq \mu\left\|\xi(T)+\frac{1}{2\sqrt{\mu}}\dot{\xi}(T)\right\|^2 - \int_0^T \lambda_1(t)\,[\text{LHS of (16)}]\,\mathrm{d}t - \int_0^T \lambda_2(t)\,[\text{LHS of (17)}]\,\mathrm{d}t \\ - \lambda_3\,[\text{LHS of (18)}] - \eta_1(\phi(0)-R_1) - \eta_2(\|\xi(0)\|^2 - R_2),\end{aligned}$$

where $\lambda_1,\lambda_2 \in C^\infty([0,T];\mathbb{R})$ and $\lambda_3,\eta_1,\eta_2 \in \mathbb{R}$. The dual problem is then written as

$$\begin{aligned}\underset{\lambda_1,\lambda_2,\lambda_3,\eta_1,\eta_2}{\text{minimize}} \quad & \max_{\phi,\xi\ \text{s.t.}\dot{\xi}(0)=0} \mathcal{L}(\xi,\phi;\lambda_1,\lambda_2,\lambda_3,\eta_1,\eta_2) \\ \text{subject to} \quad & \lambda_2(t)\geq 0, \lambda_3\geq 0, \eta_1\geq 0, \eta_2\geq 0.\end{aligned}$$

To determine the conditions on the Lagrange multipliers, it is standard to apply integration-by-parts to certain terms in $\mathcal{L}$ (with the constraint $\dot{\xi}(0)=0$) to obtain the following.

$$\begin{aligned}&\mathcal{L}(\xi,\phi;\lambda_1,\lambda_2,\lambda_3,\eta_1,\eta_2) \\ &\quad= (\lambda_3-\lambda_1(T))\phi(T) + (\lambda_1(0)-\eta_1)\phi(0) \\ &\quad+ \mu\left\|\xi(T)+\frac{1}{2\sqrt{\mu}}\dot{\xi}(T)\right\|^2 - \frac{1}{4}\lambda_1(T)\|\dot{\xi}(T)\|^2 - \frac{1}{2}\lambda_2(T)\langle\dot{\xi}(T),\xi(T)\rangle \\ &\quad- \left(\frac{\mu}{2}\lambda_3 + \frac{1}{4}(3\sqrt{\mu}\lambda_2(T)-\dot{\lambda}_2(T))\right)\|\xi(T)\|^2 + \left(\frac{1}{4}(3\sqrt{\mu}\lambda_2(0)-\dot{\lambda}_2(0))-\eta_2\right)\|\xi(0)\|^2 \\ &\quad+ \int_0^T (\dot{\lambda}_1(t)-\lambda_2(t))\phi(t)\mathrm{d}t \\ &\quad+ \int_0^T \left(\frac{1}{4}\dot{\lambda}_1(t) - \frac{3}{2}\sqrt{\mu}\lambda_1(t) + \frac{1}{2}\lambda_2(t)\right)\|\dot{\xi}(t)\|^2\mathrm{d}t + \int_0^T \left(\frac{3}{4}\sqrt{\mu}\dot{\lambda}_2(t) - \frac{1}{4}\ddot{\lambda}_2(t) - \frac{\mu}{2}\lambda_2(t)\right)\|\xi(t)\|^2\mathrm{d}t \\ &\quad+ \eta_1 R_1 + \eta_2 R_2.\end{aligned}$$

Assuming $\mathcal{L}$ is bounded, we can proceed to extract the specific conditions applicable to the multipliers. A new variable $v=\dot{\xi}$ is introduced to simplify the problem. By reducing the direct link between ξ and $\dot{\xi}$, we define a modified Lagrange function $\hat{\mathcal{L}}(\xi,v,\phi;\lambda_1,\lambda_2,\lambda_3,\eta_1,\eta_2)$. It is evident that $\max_{\xi,\phi}\mathcal{L} \leq \max_{\xi,v,\phi}\hat{\mathcal{L}}$, and the latter is sufficient to establish the desired bound.

Extracting the conditions for the multipliers in $\hat{\mathcal{L}}$ leads to the formation of a modified dual problem.

$$\begin{aligned}\underset{\lambda_1,\lambda_2,\lambda_3,\eta_1,\eta_2}{\text{minimize}} \quad & \eta_1 R_1 + \eta_2 R_2 \\ \text{subject to} \quad & \lambda_3 - \lambda_1(T) = 0, \quad \lambda_1(0) - \eta_1 = 0, \\ & \begin{bmatrix} \frac{1}{4}\lambda_1(T) & \frac{1}{4}\lambda_2(T) \\ \frac{1}{4}\lambda_2(T) & \frac{\mu}{2}\lambda_3 + \frac{1}{4}(3\sqrt{\mu}\lambda_2(T)-\dot{\lambda}_2(T)) \end{bmatrix} \succeq \begin{bmatrix} \frac{1}{4} & \frac{\sqrt{\mu}}{2} \\ \frac{\sqrt{\mu}}{2} & \mu \end{bmatrix}, \qquad (19) \\ & \frac{1}{4}(3\sqrt{\mu}\lambda_2(0)-\dot{\lambda}_2(0)) - \eta_2 \leq 0, \\ & \dot{\lambda}_1(t) - \lambda_2(t) = 0, \quad \frac{1}{4}\dot{\lambda}_1(t) - \frac{3}{2}\sqrt{\mu}\lambda_1(t) + \frac{1}{2}\lambda_2(t) \leq 0, \\ & \frac{3}{4}\sqrt{\mu}\dot{\lambda}_2(t) - \frac{1}{4}\ddot{\lambda}_2(t) - \frac{\mu}{2}\lambda_2(t) \leq 0, \qquad (20) \\ & \lambda_2(t)\geq 0, \quad \lambda_3\geq 0, \quad \eta_1\geq 0, \quad \eta_2\geq 0.\end{aligned}$$

All conditions except for the matrix semidefiniteness condition are straightforward.

This matrix condition arises from the second and third lines on the RHS of $\mathcal{L}$ ($\hat{\mathcal{L}}$). Here, the second line is the objective function, whose choice is equivalent to the tightness of the matrix condition (see Appendix A for details.)

2.3 feasible solution

Constraint (20) suggests that the ansatz $\lambda_2(t)$ might be an exponential function of t. From this assumption, a feasible solution is derived.

$$\lambda_1(t) = \mathrm{e}^{2\sqrt{\mu}(t-T)}, \quad \lambda_2(t) = 2\sqrt{\mu}\mathrm{e}^{2\sqrt{\mu}(t-T)}, \quad \lambda_3 = 1,$$
$$\eta_1 = \mathrm{e}^{-2\sqrt{\mu}T}, \quad \eta_2 = \frac{\mu}{2}\mathrm{e}^{-2\sqrt{\mu}T}.$$

This feasible solution provides an upper bound for the original problem and, consequently, the convergence rate of the ODE (10).

$$\left\| x(T) + \frac{1}{2\sqrt{\mu}}\dot{x}(T) - x^\star \right\|^2 \leq \mathrm{e}^{-2\sqrt{\mu}T}\left(\frac{R_1}{\mu} + \frac{R_2}{2}\right).$$

This finding is new in the literature, demonstrating that the convergence rate is optimal in the sense that it matches that of the TMM $\|x^{(k)} - x^\star\|^2 = \mathrm{O}((1-\sqrt{\mu/L})^{2k})$.

2.4 Extending to other analyses

The abstract cPEP* can be formulated as follows:

$$\begin{aligned} &\underset{\substack{f\in\mathcal{F}_{\mu,L},\\ x\in C^2([0,T])}}{\text{maximize}} && \mathcal{P}(f,x) \\ &\text{subject to} && (1),(11),(12). \end{aligned}$$

By selecting specific functions $\alpha(t)$ and $\beta(t)$ in ODE (1) (which determines the ODE) and a performance measure $\mathcal{P}(f,x)$ (which determines the rate), a concrete cPEP* problem is obtained. The overall strategy for solving this problem remains consistent with the aforementioned approach. Case-specific considerations are needed to achieve good convergence rates (as partially shown above) considering: (i) how the constraint (15) is relaxed, (ii) a tight matrix semidefiniteness condition (implying a performance measure), and (iii) a good feasible solution.

3 Main results via cPEP*

The main convergence theorems for ODE (1), obtained by cPEP*, are summarized in this section. The proofs are in Appendix B–Appendix D.

3.1 For strongly convex functions

The general theorem encompasses results from previous sections.

Theorem 3.1. *Let $f \in \mathcal{F}_{\mu,\infty}$ and let $x : [0,T] \to \mathbb{R}^d$ be the solution of eq. (1) with $\alpha(t) \coloneqq \sigma + 2\mu/\sigma$ and $\beta(t) \coloneqq 0$. Then, if $\sigma = 2\sqrt{\mu}$, we have for any $T \geq 0$*

$$\begin{aligned} &f(x(T)) - f^\star - \frac{\mu}{2}\|x(T)-x^\star\|^2 + \mu\left\|x(T) - x^\star + \frac{1}{2\sqrt{\mu}}\dot{x}(T)\right\|^2 \\ &\quad\leq \mathrm{e}^{-2\sqrt{\mu}T}\Big(f(x(0)) - f^\star + \frac{\mu}{2}\|x(0)-x^\star\|^2\Big). \end{aligned} \tag{21}$$

If $\sqrt{2\mu} \leq \sigma < 2\sqrt{\mu}$, we have for any $T \geq 0$

$$\begin{aligned} &f(x(T)) - f^\star + \frac{\mu}{2(4\mu-\sigma^2)}\|\sigma(x(T)-x^\star) + \dot{x}(T)\|^2 \\ &\quad\leq \frac{2\mu}{4\mu-\sigma^2}\mathrm{e}^{-\sigma T}\Big(f(x(0)) - f^\star + \frac{\mu}{2}\|x(0)-x^\star\|^2\Big). \end{aligned} \tag{22}$$

Remark 3.2. For (22), removing the restriction $\sqrt{2\mu} \leq \sigma$ is possible by setting $\alpha(t) \coloneqq 3\sigma/2$, though this increases the constant factor by four times.

When $\sigma = 2\sqrt{\mu}$, the ODE aligns with the TMM ODE (5). The first inequality in (21) indicates two types of convergence. First, the quantity $f(x(T)) - f^\star - \frac{\mu}{2}\|x(T) - x^\star\|^2$ is bounded by $\mathrm{O}\big(\mathrm{e}^{-2\sqrt{\mu}T}\big)$, in line with cPEP results. Second, the non-negativity of this quantity implies that $\|x(T) - x^\star + \dot{x}(T)/(2\sqrt{\mu})\|^2$ decreases at the rate $\mathrm{O}\big(\mathrm{e}^{-2\sqrt{\mu}T}\big)$, correlating to the new rate identified earlier.

If $\sigma < 2\sqrt{\mu}$, the convergence rate of the function is assured. The second inequality (22) suggests that by adjusting the damping α, the convergence rate of the function value can be adjusted to be as close as desired to $\mathrm{O}\big(\mathrm{e}^{-2\sqrt{\mu}T}\big)$, albeit with an increase in the constant factor. This aligns with the small $\varepsilon > 0$ in Table 2.

The inequality (22) draws inspiration from the following sense. For TMM, the convergence rate of the function value is discussed in (Van Scoy et al., 2018). However, this analysis hinges on the rate of $\|x^{(k)} - x^\star\|^2$ combined with an additional L-smoothness assumption, incorporating L as a constant in the estimate. In contrast, (22) does not require the L-smoothness assumption. This might lead to a method whose rate does not include L in the constant. Typically in ODE approaches, L-smoothness is required only post-discretization, with L primarily influencing step-size restrictions, and not influencing the constant if the convergence is linear (as detailed in (Ushiyama et al., 2023)).

The rates deduced are higher than those commonly reported in literature. For μ-strongly convex f, the AGM-SC ODE (4) is currently the fastest proven dynamical system. To compare convergence rates, time rescaling was applied so that the coefficient of the gradient term equaled 2. This adjustment results in a convergence rate of $\mathrm{O}(\mathrm{e}^{-\sqrt{2\mu}t})$ for both function value and distance to minima. Thus the convergence rate of the ODE (1) is approximately $\sqrt{2}$ times faster than the AGM-SC.

Interestingly, the LHS of (21) functions as a Lyapunov function, i.e, it does not increase with T (as detailed in Appendix F.1). Typically, Lyapunov functions are expressed as $f(x(t)) - f^\star + \Phi(t)$, where $\Phi(t)$ is a nonnegative function to show the convergence of $f(x(t)) - f^\star$ (e.g., the Lyapunov function for AGM-SC ODE (4) is defined as $f(x(t)) - f^\star + \frac{\mu}{2}\|x(t) + \frac{1}{\sqrt{2\mu}}\dot{x}(t) - x^\star\|^2$ (Wilson et al., 2021)). In this case, $\Phi(t)$ is allowed to be negative, obtaining a rapid convergence of the distance to the minima. As its price, the convergence of function values is not guaranteed.

Theorem 3.1 simplifies the analysis by setting $\beta(t)$ to 0. Further adjusting it, we can improve the constant factor (Appendix D). Also in Appendix E, we reveal new rates for the high-resolution ODEs of the heavy-ball method and AGM-SC (where $\beta(t) \neq 0$) using cPEP* , which are faster than those obtained by Lyapunov argument (Shi et al., 2022) (but not listed in Table 2 since they are not optimal.)

3.2 For both convex and strongly convex functions

The study also considers unified ODEs that exhibit sub-linear rates for convex functions and linear rates for strongly convex functions.

Theorem 3.3. *Let $f \in \mathcal{F}_{\mu,\infty}$ and let $x : [0, T] \to \mathbb{R}^d$ be the solution of eq. (1) with $\alpha(t) \coloneqq 3\sigma\coth(\sigma t)$ and $\beta(t) \coloneqq 0$. Then, if $\sigma = \sqrt{\mu}$, we have for any $T \geq 0$*

$$\begin{aligned} &f(x(T)) - f^\star - \frac{\mu}{2}\|x(T) - x^\star\|^2 + \mu\coth^2(\sqrt{\mu}T)\left\|x(T) + \frac{\tanh(\sqrt{\mu}T)}{2\sqrt{\mu}}\dot{x}(T) - x^\star\right\|^2 \\ &\quad\leq \frac{\mu}{\sinh^2(\sqrt{\mu}T)}\|x(0) - x^\star\|^2. \end{aligned} \tag{23}$$

If $\sigma < \sqrt{\mu}$, we have for any $T \geq 0$

$$f(x(T)) - f^\star + \frac{\mu}{4(\mu - \sigma^2)}\|2\sigma\coth(\sigma T)(x(T) - x^\star) + \dot{x}(T)\|^2 \leq \frac{\mu}{\mu - \sigma^2}\frac{\sigma^2}{\sinh^2(\sigma T)}\|x(0) - x^\star\|^2.$$

When $\sigma = \sqrt{\mu}$, the ODE aligns with the ITEM ODE (7). Similar to Theorem 3.1, from (23), the convergence result of $f(x(T)) - f^\star - \frac{\mu}{2}\|x(T) - x^\star\|^2$, as obtained using cPEP, is recoverable. A new result concerning the distance from the optimal solution is also presented

$$\left\|x(T) + \frac{\tanh(\sqrt{\mu}T)}{2\sqrt{\mu}}\dot{x}(T) - x^\star\right\|^2 \leq \frac{1}{\cosh^2(\sqrt{\mu}T)}\|x(0) - x^\star\|^2.$$

In the limit where $\mu \to 0$, the damping term in the equation changes to $\alpha(t) = 3/t$, with (23) being

$$f(x(T)) - f^\star + \frac{1}{4}\left\|\frac{2}{T}(x(T) - x^\star) + \dot{x}(T)\right\|^2 \leq \frac{1}{T^2}\|x(0) - x^\star\|^2,$$

which aligns with the result of the OGM ODE (3), as presented in Theorem 3.5. Moreover, the LHS of (23) serves as a Lyapunov function (as elaborated in Appendix F.2). This alignment establishes an exact correspondence between the ITEM and its corresponding ODE model.

When $\sigma < \sqrt{\mu}$, the convergence rate of the function value is assured. Similar to TMM, adjusting the damping α allows for an arbitrarily close approximation of the convergence rate concerning $f - f^\star$ to $\mathrm{O}\big(\mu \sinh^{-2}(\sqrt{\mu}T)\big)$, albeit with an increase in the constant factor. The hyperbolic convergence rate implies $\sigma^2/\sinh^2(\sigma T) \leq \min\big\{1/T^2, \mathrm{e}^{-2\sigma T}\big\}$, which is faster than the (time-rescaled) unified AGM ODE.

For the ITEM-G ODE (9), we discuss the convergence rate of the gradient norm. The result aligns with that obtained using cPEP, as documented in Appendix of (Kim & Yang, 2023b).

Theorem 3.4. *Let $f \in \mathcal{F}_{0,\infty}$, and let $x : [0, S) \to \mathbb{R}^d$ be the solution of eq. (1) with $\alpha(t) \coloneqq 3\sqrt{\mu}\coth(\sqrt{\mu}(S-t))$ and $\beta(t) \coloneqq 0$ where $S > 0$. Then, for any $0 \leq T < S$, we have*

$$\left\| \frac{\dot{x}(T)}{\sinh^2(\sqrt{\mu}(S-T))} \right\|^2 \leq \frac{4}{\sinh^2(\sqrt{\mu}S)} \Big(f(x(0)) - f^\star + \frac{\mu}{2}\|x(0) - x(T)\|^2 \Big).$$

This finding implies that $\lim_{T\to S} \dot{x}(T) = 0$, and $\|\ddot{x}(T)\|^2 \leq (\text{const.})/\sinh^2(\sqrt{\mu}S)$. In the limit as $t \to S$ for the ODE (9), we have $\ddot{x}(S) = \nabla f(x(S))$, and

$$\|\nabla f(x(S))\|^2 \leq \frac{4\mu}{\sinh^2(\sqrt{\mu}S)} \Big(f(x(0)) - f^\star + \frac{\mu}{2}\|x(0) - x(T)\|^2 \Big).$$

Furthermore, taking the limit $\mu \to 0$ recovers Theorem 3.7. Note that ITEM-G has so far only been explored at the ODE level, with no corresponding discrete method yet developed. These results may suggest the potential existence of an efficient method for minimizing the gradient norm through the discretization of the ITEM-G ODE.

3.3 For convex functions

Classical results of the Lyapunov function approach can be re-established using cPEP* for all standard quantities.

By setting $\alpha(t) = 3/t$, ODE (1) becomes the OGM ODE (3), exhibiting a $\sqrt{2}$ times faster time scale than the AGM ODE (2). Consequently, it achieves better constants than the well-known AGM ODE, but this advantage is only due to the difference in the time scale.

Theorem 3.5. *Let $f \in \mathcal{F}_{0,L}$ ($L = \infty$ allowed) and let $x : [0, T] \to \mathbb{R}^d$ be the solution of eq. (1) with $\alpha(t) \coloneqq a/t$ and $\beta(t) \coloneqq b$ where $a \geq 3$ and $b \geq 0$. Then, for any $T \geq 0$, we have*

$$f(x(T) + b\dot{x}(T)) - f^\star \leq \frac{a-1}{2T^2}\|x(0) - x^\star\|^2, \tag{24}$$

and

$$\int_0^T \left(\frac{3a-2}{a(a-1)} bt^2 + \frac{t}{aL} \right) \|\nabla f(x(t) + \beta(t)\dot{x}(t))\|^2 \mathrm{d}t \leq \|x(0) - x^\star\|^2.$$

This immediately implies

$$\min_{0\leq t\leq T} \|\nabla f(x(t) + \beta(t)\dot{x}(t))\|^2 \leq \left(\frac{3a-2}{3a(a-1)} bT^3 + \frac{2T^2}{aL} \right)^{-1} \|x(0) - x^\star\|^2.$$

Moreover, if $a > 3$, we have

$$\int_0^T t(f(x(t) + b\dot{x}(t)) - f^\star)\mathrm{d}t \leq \frac{(a-1)^2}{4(a-3)}\|x(0) - x^\star\|^2.$$

This and (24) immediately imply $f(x(T) + b\dot{x}(T)) - f^\star = \mathrm{o}\big(1/T^2\big)$.

Remark 3.6. For the first and second inequality, we have the best rate at $a = 3$, while for the last inequality the best is achieved at $a = 5$.

For OGM-G ODE (8), we have the same convergence result as (Suh et al., 2022).

Theorem 3.7. *Let $f \in \mathcal{F}_{0,\infty}$, and let $x : [0, S) \to \mathbb{R}^d$ be the solution of eq. (1) with $\alpha(t) \coloneqq 3/(S-t)$ and $\beta(t) \coloneqq 0$ where $S > 0$. Then, for any $0 \le T < S$, we have*

$$\left\| \frac{\dot{x}(T)}{S-T} \right\|^2 \le \frac{4}{S^2}(f(x(0)) - f^\star).$$

Taking the limit of $T \to S$, we have

$$\|\nabla f(x(S))\|^2 \le \frac{4}{S^2}(f(x(0)) - f^\star).$$

4 Discussions and conclusions

How we compare cPEP and cPEP*. The two frameworks have their own advantage and disadvantages, which are somewhat complimentary. The framework of cPEP offers clarity and a unified analysis approach using the IDE form, somewhat paralleling discrete PEP. cPEP* has a less direct correspondence with discrete PEP due to the absence of the SDP-type argument and requires experience and trial and error in (i), (ii), and (iii), as indicated in 2. Instead, there is much more room in cPEP* to optimize discussions of fully utilizing L-smoothness and μ-strong convexity and adapt to a variety of performance measures. This gives rise to (near-)optimal convergence rates for standard performance measures: function values, distance to the minimum, and gradient norm. In addition, the second-order ODE (1) can handle ODEs that are difficult to express in IDEs, including those with shifted gradients.

Obtaining better discrete methods from the optimal ODE models. The next research goal is to develop new discrete methods from optimal ODEs that outperform current methods. This task is challenging due to the faster time scales of ODEs in this study compared to classic models such as AGM ODE, which might complicate stable discretization. The feasibility of stable discretization for these faster ODE models remains an open question.

Mysterious rates for L-smooth functions. For a gradient flow $\dot{x} = -\nabla f(x)$ of μ-strongly convex and L-smooth functions f, an unusual bound is known: $\|x(T) - x^\star\|^2 \le \mathrm{e}^{-\frac{2\mu L}{\mu+L}T}\|x(0) - x^\star\|^2$ (Wilson, 2018), which worsens as L decreases. This paradox is not resolved by cPEP*. For example, the rate for the ODE considered in Theorem 3.1 is up to $\mathrm{O}\left(\mathrm{e}^{-\sqrt{\mu(4-2\mu/L)}T}\right)$ (see Appendix B.2.6). It is still unclear whether this deterioration can be avoided, leaving it as an open question.

References

Attouch, H. and Cabot, A. Asymptotic stabilization of inertial gradient dynamics with time-dependent viscosity. *J. Differential Equations*, 263(9):5412–5458, 2017. doi: 10.1016/j.jde.2017.06.024.

Attouch, H., Peypouquet, J., and Redont, P. Fast convex optimization via inertial dynamics with Hessian driven damping. *J. Differential Equations*, 261(10):5734–5783, 2016. doi: 10.1016/j.jde.2016.08.020.

Attouch, H., Chbani, Z., Peypouquet, J., and Redont, P. Fast convergence of inertial dynamics and algorithms with asymptotic vanishing viscosity. *Math. Program.*, 168(1-2, Ser. B):123–175, 2018a. doi: 10.1007/s10107-016-0992-8.

Attouch, H., Chbani, Z., and Riahi, H. Combining fast inertial dynamics for convex optimization with Tikhonov regularization. *J. Math. Anal. Appl.*, 457(2):1065–1094, 2018b. doi: 10.1016/j.jmaa.2016.12.017.

Attouch, H., Chbani, Z., and Riahi, H. Rate of convergence of the Nesterov accelerated gradient method in the subcritical case $\alpha \le 3$. *ESAIM Control Optim. Calc. Var.*, 25:Paper No. 2, 34, 2019. doi: 10.1051/cocv/2017083.

Attouch, H., Balhag, A., Chbani, Z., and Riahi, H. Damped inertial dynamics with vanishing Tikhonov regularization: strong asymptotic convergence towards the minimum norm solution. *J. Differential Equations*, 311:29–58, 2022a. doi: 10.1016/j.jde.2021.12.005.

Attouch, H., Balhag, A., Chbani, Z., and Riahi, H. Fast convex optimization via inertial dynamics combining viscous and Hessian-driven damping with time rescaling. *Evol. Equ. Control Theory*, 11(2):487–514, 2022b. doi: 10.3934/eect.2021010.

Attouch, H., Chbani, Z., Fadili, J., and Riahi, H. First-order optimization algorithms via inertial systems with Hessian driven damping. *Math. Program.*, 193(1, Ser. A):113–155, 2022c. doi: 10.1007/s10107-020-01591-1.

Aujol, J. and Dossal, C. Optimal rate of convergence of an ODE associated to the fast gradient descent schemes for $b > 0$. hal-01547251v2, 2017.

Bansal, N. and Gupta, A. Potential-function proofs for gradient methods. *Theory Comput.*, 15:Paper No. 4, 32, 2019. doi: 10.4086/toc.2019.v015a004.

Boţ, R. I., Csetnek, E. R., and László, S. C. Tikhonov regularization of a second order dynamical system with Hessian driven damping. *Math. Program.*, 189(1-2, Ser. B):151–186, 2021. doi: 10.1007/s10107-020-01528-8.

Chen, S., Shi, B., and Yuan, Y.-x. Gradient norm minimization of Nesterov acceleration: $o(1/k^3)$. *arXiv preprint arXiv:2209.08862*, 2022.

Chen, S., Shi, B., and Yuan, Y.-x. On underdamped Nesterov's acceleration. *arXiv preprint arXiv:2304.14642*, 2023.

Diakonikolas, J. and Orecchia, L. The approximate duality gap technique: a unified theory of first-order methods. *SIAM J. Optim.*, 29(1):660–689, 2019. doi: 10.1137/18M1172314.

Drori, Y. The exact information-based complexity of smooth convex minimization. *J. Complexity*, 39:1–16, 2017. doi: 10.1016/j.jco.2016.11.001.

Drori, Y. and Taylor, A. On the oracle complexity of smooth strongly convex minimization. *J. Complexity*, 68:101590, 2022. doi: 10.1016/j.jco.2021.101590.

Drori, Y. and Teboulle, M. Performance of first-order methods for smooth convex minimization: a novel approach. *Math. Program.*, 145(1-2, Ser. A):451–482, 2014. doi: 10.1007/s10107-013-0653-0.

Fazlyab, M., Ribeiro, A., Morari, M., and Preciado, V. M. Analysis of optimization algorithms via integral quadratic constraints: nonstrongly convex problems. *SIAM J. Optim.*, 28(3):2654–2689, 2018. doi: 10.1137/17M1136845.

Foret, P., Kleiner, A., Mobahi, H., and Neyshabur, B. Sharpness-aware minimization for efficiently improving generalization. In *International Conference on Learning Representations*, 2021.

Goujaud, B., Moucer, C., Glineur, F., Hendrickx, J., Taylor, A., and Dieuleveut, A. PEPit: computer-assisted worst-case analyses of first-order optimization methods in Python. *arXiv preprint arXiv:2201.04040*, 2022.

Kim, D. and Fessler, J. A. Optimized first-order methods for smooth convex minimization. *Math. Program.*, 159(1-2, Ser. A):81–107, 2016. doi: 10.1007/s10107-015-0949-3.

Kim, D. and Fessler, J. A. Optimizing the efficiency of first-order methods for decreasing the gradient of smooth convex functions. *J. Optim. Theory Appl.*, 188(1):192–219, 2021. doi: 10.1007/s10957-020-01770-2.

Kim, J. and Yang, I. Unifying Nesterov's accelerated gradient methods for convex and strongly convex objective functions. In *International Conference on Machine Learning*, pp. 16897–16954. PMLR, 2023a.

Kim, J. and Yang, I. Convergence analysis of ODE models for accelerated first-order methods via positive semidefinite kernels. In *Advances in Neural Information Processing Systems*, volume 37, 2023b.

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.

Lessard, L., Recht, B., and Packard, A. Analysis and design of optimization algorithms via integral quadratic constraints. *SIAM J. Optim.*, 26(1):57–95, 2016. doi: 10.1137/15M1009597.

Lu, H. An $O(s^r)$-resolution ODE framework for understanding discrete-time algorithms and applications to the linear convergence of minimax problems. *Math. Program.*, 194(1-2, Ser. A):1061–1112, 2022. doi: 10.1007/s10107-021-01669-4.

Maskan, H., Zygalakis, K. C., and Yurtsever, A. A variational perspective on high-resolution ODEs. In *Advances in Neural Information Processing Systems*, volume 37, 2023.

May, R. Asymptotic for a second-order evolution equation with convex potential and vanishing damping term. *Turkish J. Math.*, 41(3):681–685, 2017. doi: 10.3906/mat-1512-28.

Moucer, C., Taylor, A., and Bach, F. A systematic approach to Lyapunov analyses of continuous-time models in convex optimization. *SIAM J. Optim.*, 33(3):1558–1586, 2023. doi: 10.1137/22M1498486.

Nesterov, Y. *Lectures on convex optimization*, volume 137 of *Springer Optimization and Its Applications*. Springer, Cham, 2018. ISBN 978-3-319-91577-7; 978-3-319-91578-4. doi: 10.1007/978-3-319-91578-4.

Nesterov, Y., Gasnikov, A., Guminov, S., and Dvurechensky, P. Primal-dual accelerated gradient methods with small-dimensional relaxation oracle. *Optim. Methods Softw.*, 36(4):773–810, 2021. doi: 10.1080/10556788.2020.1731747.

Nesterov, Y. E. A method for solving the convex programming problem with convergence rate $O(1/k^2)$. *Dokl. Akad. Nauk SSSR*, 269(3):543–547, 1983.

Polyak, B. T. Some methods of speeding up the convergence of iterative methods. *USSR Comput. Math. Math. Phys.*, 4(5): 1–17, 1964. doi: https://doi.org/10.1016/0041-5553(64)90137-5.

Sanz Serna, J. M. and Zygalakis, K. C. The connections between Lyapunov functions for some optimization algorithms and differential equations. *SIAM J. Numer. Anal.*, 59(3):1542–1565, 2021. doi: 10.1137/20M1364138.

Shi, B., Du, S. S., Jordan, M. I., and Su, W. J. Understanding the acceleration phenomenon via high-resolution differential equations. *Math. Program.*, 195(1-2):79–148, 2022. doi: 10.1007/s10107-021-01681-8.

Su, W., Boyd, S., and Candès, E. J. A differential equation for modeling Nesterov's accelerated gradient method: theory and insights. In *Advances in Neural Information Processing Systems*, volume 27, 2014.

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, 2022.

Sun, B., George, J., and Kia, S. High-resolution modeling of the fastest first-order optimization method for strongly convex functions. In *2020 59th IEEE Conference on Decision and Control (CDC)*, pp. 4237–4242, 2020. doi: 10.1109/CDC42340.2020.9304444.

Taylor, A. and Bach, F. Stochastic first-order methods: non-asymptotic and computer-aided analyses via potential functions. In *Proceedings of the Thirty-Second Conference on Learning Theory*, volume 99 of *Proceedings of Machine Learning Research*, pp. 2934–2992. PMLR, 25–28 Jun 2019.

Taylor, A. and Drori, Y. An optimal gradient method for smooth strongly convex minimization. *Math. Program.*, pp. 1–38, 2022. doi: 10.1007/s10107-022-01839-y.

Taylor, A., Van Scoy, B., and Lessard, L. Lyapunov functions for first-order methods: Tight automated convergence guarantees. In *International Conference on Machine Learning*. PMLR, 2018.

Taylor, A. B., Hendrickx, J. M., and Glineur, F. Smooth strongly convex interpolation and exact worst-case performance of first-order methods. *Math. Program.*, 161(1-2):307–345, 2017a. doi: 10.1007/s10107-016-1009-3.

Taylor, A. B., Hendrickx, J. M., and Glineur, F. Performance estimation toolbox (PESTO): Automated worst-case analysis of first-order optimization methods. In *2017 IEEE 56th Annual Conference on Decision and Control (CDC)*, pp. 1278–1283, 2017b. doi: 10.1109/CDC.2017.8263832.

Ushiyama, K., Sato, S., and Matsuo, T. Essential convergence rate of ordinary differential equations appearing in optimization. *JSIAM Lett.*, 14:119–122, 2022.

Ushiyama, K., Sato, S., and Matsuo, T. A unified discretization framework for differential equation approach with Lyapunov arguments for convex optimization. In *Advances in Neural Information Processing Systems*, volume 37, 2023.

Van Scoy, B., Freeman, R. A., and Lynch, K. M. The fastest known globally convergent first-order method for minimizing strongly convex functions. *IEEE Control Syst. Lett.*, 2(1):49–54, 2018.

Wibisono, A., Wilson, A. C., and Jordan, M. I. A variational perspective on accelerated methods in optimization. *Proc. Natl. Acad. Sci. USA*, 113(47):E7351–E7358, 2016. doi: 10.1073/pnas.1614734113.

Wilson, A. *Lyapunov arguments in optimization*. PhD thesis, University of California, Berkeley, 2018.

Wilson, A. C., Recht, B., and Jordan, M. I. A Lyapunov analysis of accelerated methods in optimization. *J. Mach. Learn. Res.*, 22(113):1–34, 2021.

A For better rates: optimizing discussions in cPEP*

The procedure of cPEP* is as described in Section 2. It is quite descriptive and can handle various cases, but slight change of discussion can lead to different results (convergence rates), and some careful consideration there (possibly by some trial and error) is essential part of the overall framework—recall the points (i)—(iii) in Section 2.4. Here we illustrate it by taking the discussion in Section 2 as an example.

The main adjustment in Section 2 lies in (ii). There the magic starts in Section 2.2, where we considered the (modified) dual problem. There, the objective function is

$$\mu\left\|\xi(T)+\frac{1}{2\sqrt{\mu}}\dot{\xi}(T)\right\|^2, \qquad \text{where } \xi := x-x^\star.$$

One might feel it is weird, and

$$\mu\|\xi(T)\|^2=\mu\|x(T)-x^\star\|^2 \tag{25}$$

is more natural, but this is exactly where the magic lies.

Observe that the objective function (performance measure) is set at the very start of the procedure, and left untouched until this dual problem. This means that we can still change it (so that we can reach a better result) without destroying the overall discussion.

Let us see how the story would go if we choose the measure (25). This affects the matrix constraint (19). It is exactly equivalent to

$$\begin{aligned}&\mu\left\|\xi(T)+\frac{1}{2\sqrt{\mu}}\dot{\xi}(T)\right\|^2\\&-\frac{1}{4}\lambda_1(T)\|\dot{\xi}(T)\|^2-\frac{1}{2}\lambda_2(T)\langle\dot{\xi}(T),\xi(T)\rangle\\&-\left(\frac{\mu}{2}\lambda_3+\frac{1}{4}(3\sqrt{\mu}\lambda_2(T)-\dot{\lambda}_2(T))\right)\|\xi(T)\|^2\le 0,\end{aligned}$$

which gives a bound for the lines 2–4 of the RHS of the Lagrange function. Observe that the objective function here corresponds to the constant matrix in the RHS of (19). If we change the objective function to (25), the matrix becomes

$$\begin{bmatrix}0&0\\0&\mu\end{bmatrix} \qquad \left(\preceq\begin{bmatrix}\frac{1}{4}\lambda_1(T) & \frac{1}{4}\lambda_2(T)\\ \frac{1}{4}\lambda_2(T) & \frac{\mu}{2}\lambda_3+\frac{1}{4}(3\sqrt{\mu}\lambda_2(T)-\dot{\lambda}_2(T))\end{bmatrix}\right).$$

Accordingly, we reach a weaker feasible solution:

$$\lambda_1(t)=\frac{4}{3}\mathrm{e}^{\sqrt{\mu}(t-T)},\quad \lambda_2(t)=\frac{4}{3}\sqrt{\mu}\mathrm{e}^{\sqrt{\mu}(t-T)},\quad \lambda_3=\frac{3}{4},$$
$$\eta_1=\frac{4}{3}\mathrm{e}^{-\sqrt{\mu}T},\quad \eta_2=\frac{2\mu}{3}\mathrm{e}^{-\sqrt{\mu}T},$$

and we have a convergence estimate

$$\|x(T)-x^\star\|^2\le \mathrm{e}^{-\sqrt{\mu}T}\left(\frac{4R_1}{3\mu}+\frac{2R_2}{3}\right),$$

which is weaker than the rate obtained in the main body, $\sim \mathrm{e}^{-2\sqrt{\mu}T}$.

The above observation gives us the following lessons. The setting of the performance measure only affects the discussion in the last stage of the procedure, i.e., when we finally establish a feasible solution. This, in turn, strongly affects the resulting rate. Thus it is worth taking a breath at this point and consider if we can optimize the overall discussion by adjusting the performance measure for making the resulting rate the best possible. Similar adjustment has been done for the main theorems in Section 3.

Other points, (i) and (iii), do not matter so much in Section 2 (the example is chosen from this reason), but should be cared in more general cases. For the point (i), how we relax the constraint (15), in this study we prepared two options: one for function values and distances to minimum (Appendix B), and the other for gradient norms (Appendix C). For the point (iii), the feasible solutions, in this study we show the best solutions we found. Note that in the general theory, we sometimes adjust (ii) and (iii) at the same time (see the proof in Appendix B and Appendix C).

B cPEP* for function values and distances to minimum

In this section, we prove Theorems 3.1, 3.3 and 3.5, and show an additional result. Since the proofs of these theorems involve similar arguments, we first show the common part under a generalized objective function $\mathcal{P}$. A relaxed problem is derived in Appendix B.1 and the Lagrange function is simplified in the first half of Appendix B.2. Then, in Appendices B.2.1 to B.2.6, we show the case specific discussions. Specifically, Appendices B.2.1 to B.2.3 show the three inequalities in Theorem 3.5, Appendix B.2.4 shows the second halves of Theorems 3.1 and 3.3, Appendix B.2.5 shows the first halves of Theorems 3.1 and 3.3, and the case of finite L is considered in Appendix B.2.6.

We consider the following continuous-time PEP:

$$\begin{aligned}
\underset{\substack{f\in\mathcal{F}_{\mu,L},\\ x\in C^2([0,T];\mathbb{R}^d)}}{\text{maximize}} \quad & \mathcal{P}(x,\alpha,\beta),\\
\text{subject to} \quad & \ddot{x}(t)+\alpha(t)\dot{x}(t)+2\nabla f(x(t)+\beta(t)\dot{x}(t))=0,\\
& f(x(0))-f^\star\le R_1,\\
& \|x(0)-x^\star\|^2\le R_2,\quad \dot{x}(0)=0.
\end{aligned}$$

By introducing $\xi(t)\coloneqq x(t)-x^\star$, we rewrite the problem as

$$\begin{aligned}
\underset{\substack{f\in\mathcal{F}_{\mu,L},\\ \xi\in C^2([0,T];\mathbb{R}^d),\\ \phi\in C^1([0,T];\mathbb{R}),\\ \gamma\in C([0,T];\mathbb{R}^d)}}{\text{maximize}} \quad & \mathcal{P}(\xi+x^\star,\alpha,\beta),\\
\text{subject to} \quad & \phi(t)=f\Big(\xi+\beta\dot{\xi}+x^\star\Big)-f^\star,\\
& \gamma(t)=\nabla f\Big(\xi+\beta\dot{\xi}+x^\star\Big),\\
& \ddot{\xi}(t)+\alpha(t)\dot{\xi}(t)+2\nabla f\Big(\xi(t)+x^\star+\beta(t)\dot{\xi}(t)\Big)=0,\\
& \phi(0)\le R_1,\\
& \|\xi(0)\|^2\le R_2,\quad \dot{\xi}(0)=0.
\end{aligned}$$

B.1 Relaxing the problem

Throughout the appendix, ∂_t denotes the temporal differential operator. Hereafter, we sometimes omit the dependence of $\xi(t),\phi(t),\gamma(t)$ on t for simplicity. Since the most problematic part of the problem is the constraint $f\in\mathcal{F}_{\mu,L}$, in this section, we derive several inequalities from the constraint and relax the problem. The relaxation below corresponds to that in Section 2.1. This part corresponds to (i) in Section 2.4: for the cases considered in this section, the common relaxation below is enough to derive the optimal rate.

From the chain rule, we see

$$\begin{aligned}
\partial_t f(\xi+\beta\dot{\xi}+x^\star) &= \Big\langle \nabla f(\xi+\beta\dot{\xi}+x^\star), \partial_t(\xi+\beta\dot{\xi}+x^\star)\Big\rangle\\
&= \Big\langle \nabla f(\xi+\beta\dot{\xi}+x^\star), (1+\dot{\beta})\dot{\xi}+\beta\ddot{\xi}\Big\rangle\\
&= \Big\langle \nabla f(\xi+\beta\dot{\xi}+x^\star), (1+\dot{\beta})\dot{\xi}-\beta(\alpha\dot{\xi}+2\nabla f(\xi+\beta\dot{\xi}+x^\star))\Big\rangle\\
&= \Big(1+\dot{\beta}-\beta\alpha\Big)\Big\langle \nabla f(\xi+\beta\dot{\xi}+x^\star),\dot{\xi}\Big\rangle-2\beta\Big\|\nabla f(\xi+\beta\dot{\xi}+x^\star)\Big\|^2,
\end{aligned}$$

which implies

$$\dot{\phi}=\Big(1+\dot{\beta}-\beta\alpha\Big)\Big\langle\gamma,\dot{\xi}\Big\rangle-2\beta\|\gamma\|^2.$$

By using the ODE (1), we have

$$\dot{\phi}=-\frac{1+\dot{\beta}-\beta\alpha}{2}\Big\langle\ddot{\xi}+\alpha\dot{\xi},\dot{\xi}\Big\rangle-\frac{\beta}{2}\Big\|\ddot{\xi}+\alpha\dot{\xi}\Big\|^2$$

which implies

$$\dot{\phi} + \frac{1+\dot{\beta}}{2}\alpha\left\|\dot{\xi}\right\|^2 + \frac{1+\dot{\beta}+\beta\alpha}{2}\left\langle\ddot{\xi},\dot{\xi}\right\rangle + \frac{\beta}{2}\left\|\ddot{\xi}\right\|^2 = 0. \tag{26}$$

Since we assume the objective function f is L-smooth and μ-strongly convex,

$$f(u)-f(v)-\langle\nabla f(v),u-v\rangle \geq \frac{L}{2(L-\mu)}\left(\frac{1}{L}\|\nabla f(u)-\nabla f(v)\|^2 + \mu\|u-v\|^2 - 2\frac{\mu}{L}\langle\nabla f(v)-\nabla f(u),v-u\rangle\right) \tag{27}$$

holds for all $u,v\in\mathbb{R}^d$. By substituting $u=x^\star$ and $v=\xi+\beta\dot{\xi}+x^\star$ (this choice corresponds to the case $(\phi',\gamma',\xi')=(0,0,0)$, $(\phi'',\gamma'',\xi'')\in\mathcal{G}(T)$ in Section 2.1), we have

$$\phi - \left\langle\gamma,\xi+\beta\dot{\xi}\right\rangle + \frac{L}{2(L-\mu)}\left(\frac{1}{L}\|\gamma\|^2 + \mu\left\|\xi+\beta\dot{\xi}\right\|^2 - 2\frac{\mu}{L}\left\langle\gamma,\xi+\beta\dot{\xi}\right\rangle\right) \leq 0,$$

which can be simplified as

$$\phi - \frac{L}{L-\mu}\left\langle\gamma,\xi+\beta\dot{\xi}\right\rangle + \frac{1}{2(L-\mu)}\|\gamma\|^2 + \frac{L\mu}{2(L-\mu)}\left\|\xi+\beta\dot{\xi}\right\|^2 \leq 0,$$

By using the ODE (1), we have

$$\phi + \frac{L}{2(L-\mu)}\left\langle\ddot{\xi}+\alpha\dot{\xi},\xi+\beta\dot{\xi}\right\rangle + \frac{1}{8(L-\mu)}\left\|\ddot{\xi}+\alpha\dot{\xi}\right\|^2 + \frac{L\mu}{2(L-\mu)}\left\|\xi+\beta\dot{\xi}\right\|^2 \leq 0,$$

which implies

$$\begin{aligned}\phi + \frac{4\alpha\beta L+\alpha^2+4L\mu\beta^2}{8(L-\mu)}\left\|\dot{\xi}\right\|^2 + \frac{2\beta L+\alpha}{4(L-\mu)}\left\langle\ddot{\xi},\dot{\xi}\right\rangle + \frac{(\alpha+2\beta\mu)L}{2(L-\mu)}\left\langle\dot{\xi},\xi\right\rangle \\ + \frac{L\mu}{2(L-\mu)}\|\xi\|^2 + \frac{L}{2(L-\mu)}\left\langle\ddot{\xi},\xi\right\rangle + \frac{1}{8(L-\mu)}\left\|\ddot{\xi}\right\|^2 \leq 0.\end{aligned} \tag{28}$$

Finally, by substituting $u=\xi(T)+x^\star$ and $v=x^\star$ (this choice corresponds to the case $(\phi',\gamma',\xi')=(\phi(T),\gamma(T),\xi(T))$, $(\phi'',\gamma'',\xi'')=(0,0,0)$ in Section 2.1) to (27), we have

$$-\phi(T) + \frac{L}{2(L-\mu)}\left(\frac{1}{L}\|\gamma(T)\|^2 + \mu\|\xi(T)\|^2 - 2\frac{\mu}{L}\langle\gamma(T),\xi(T)\rangle\right) \leq 0. \tag{29}$$

By using the equality (26) and the inequalities (28) and (29), we can relax the problem as

$$\begin{aligned}\underset{\substack{\xi\in C^2([0,T];\mathbb{R}^d),\\ \phi\in C^1([0,T];\mathbb{R}),\\ \gamma(T)\in\mathbb{R}^d}}{\text{maximize}} \quad & \mathcal{P}(\xi+x^\star,\alpha,\beta), \\ \text{subject to} \quad & \text{equality (26)}, \\ & \text{inequality (28)}, \\ & \text{inequality (29)}, \\ & \phi(0)\leq R_1, \\ & \|\xi(0)\|^2 \leq R_2, \quad \dot{\xi}(0)=0.\end{aligned}$$

B.2 Lagrange dual problem

We define the Lagrange function as

$$\begin{aligned}\mathcal{L}(\xi,\phi,\gamma(T);\lambda_1,\lambda_2,\lambda_3,\eta_1,\eta_2) \coloneqq \mathcal{P}(\xi+x^\star,\alpha,\beta) - \int_0^T \lambda_1(t)\,[\text{LHS of (26)}]\,\mathrm{d}t - \int_0^T \lambda_2(t)\,[\text{LHS of (28)}]\,\mathrm{d}t \\ - \lambda_3\,[\text{LHS of (29)}] - \eta_1(\phi(0)-R_1) - \eta_2(\|\xi(0)\|^2 - R_2),\end{aligned}$$

where $\lambda_1, \lambda_2 \in C^\infty([0,T];\mathbb{R})$ and $\lambda_3, \eta_1, \eta_2 \in \mathbb{R}$. The Lagrange dual problem reads

$$\begin{aligned}&\underset{\lambda_1,\lambda_2,\lambda_3,\eta_1,\eta_2}{\text{minimize}} \ \underset{\xi \text{ s.t. } \dot\xi(0)=0,\phi,\gamma(T)}{\text{maximize}} \quad \mathcal{L}(\xi,\phi,\gamma(T);\lambda_1,\lambda_2,\lambda_3,\eta_1,\eta_2),\\ &\text{subject to} \quad \lambda_2(t)\geq 0, \lambda_3\geq 0, \eta_1\geq 0, \eta_2\geq 0.\end{aligned}$$

To simplify the dual problem, we try to eliminate the primal variables $x, \phi, \gamma(T)$.

First, we consider the term $\int_0^T \lambda_1(t)\,[\text{LHS of (26)}]\,\mathrm{d}t$. Using integration-by-parts and $\dot\xi(0)=0$, we see

$$\begin{aligned}\int_0^T \lambda_1\,[\text{LHS of (26)}]\,\mathrm{d}t &= \int_0^T \lambda_1\left(\dot\phi + \frac{1+\dot\beta}{2}\alpha\left\|\dot\xi\right\|^2 + \frac{1+\dot\beta+\beta\alpha}{2}\left\langle\ddot\xi,\dot\xi\right\rangle + \frac{\beta}{2}\left\|\ddot\xi\right\|^2\right)\mathrm{d}t\\ &= [\lambda_1(t)\phi(t)]_0^T - \int_0^T \dot\lambda_1\phi\,\mathrm{d}t + \int_0^T \lambda_1\frac{1+\dot\beta}{2}\alpha\left\|\dot\xi\right\|^2\mathrm{d}t + \int_0^T \frac{\lambda_1\beta}{2}\left\|\ddot\xi\right\|^2\mathrm{d}t\\ &\quad + \frac{1}{4}\int_0^T \lambda_1\Big(1+\dot\beta+\beta\alpha\Big)\partial_t\left\|\dot\xi\right\|^2\mathrm{d}t\\ &= \lambda_1(T)\phi(T) - \lambda_1(0)\phi(0) - \int_0^T \dot\lambda_1\phi\,\mathrm{d}t + \int_0^T \lambda_1\frac{1+\dot\beta}{2}\alpha\left\|\dot\xi\right\|^2\mathrm{d}t + \int_0^T \frac{\lambda_1\beta}{2}\left\|\ddot\xi\right\|^2\mathrm{d}t\\ &\quad + \frac{1}{4}\left[\lambda_1(t)\Big(1+\dot\beta(t)+\beta(t)\alpha(t)\Big)\left\|\dot\xi(t)\right\|^2\right]_0^T - \frac{1}{4}\int_0^T \partial_t\Big(\lambda_1\Big(1+\dot\beta+\beta\alpha\Big)\Big)\left\|\dot\xi\right\|^2\mathrm{d}t\\ &= \lambda_1(T)\phi(T) - \lambda_1(0)\phi(0) + \frac{1}{4}\lambda_1(T)\Big(1+\dot\beta(T)+\beta(T)\alpha(T)\Big)\left\|\dot\xi(T)\right\|^2 - \int_0^T \dot\lambda_1\phi\,\mathrm{d}t\\ &\quad + \frac{1}{2}\int_0^T\left(\lambda_1\Big(1+\dot\beta\Big)\alpha - \frac{1}{2}\partial_t\Big(\lambda_1\Big(1+\dot\beta+\beta\alpha\Big)\Big)\right)\left\|\dot\xi\right\|^2\mathrm{d}t + \int_0^T \frac{\lambda_1\beta}{2}\left\|\ddot\xi\right\|^2\mathrm{d}t.\end{aligned}$$

Second, we consider the term $\int_0^T \lambda_2(t)\,[\text{LHS of (28)}]\,\mathrm{d}t$. Using integration-by-parts and $\dot\xi(0)=0$, we see

$$\begin{aligned}\int_0^T \lambda_2\,[\text{LHS of (28)}]\,\mathrm{d}t &= \int_0^T \lambda_2\left(\phi + \frac{4\alpha\beta L+\alpha^2+4L\mu\beta^2}{8(L-\mu)}\left\|\dot\xi\right\|^2 + \frac{2\beta L+\alpha}{4(L-\mu)}\left\langle\ddot\xi,\dot\xi\right\rangle + \frac{(\alpha+2\beta\mu)L}{2(L-\mu)}\left\langle\dot\xi,\xi\right\rangle\right.\\ &\qquad \left. + \frac{L\mu}{2(L-\mu)}\|\xi\|^2 + \frac{L}{2(L-\mu)}\left\langle\ddot\xi,\xi\right\rangle + \frac{1}{8(L-\mu)}\left\|\ddot\xi\right\|^2\right)\mathrm{d}t\\ &= \int_0^T \lambda_2\phi\,\mathrm{d}t + \int_0^T \lambda_2\frac{4\alpha\beta L+\alpha^2+4L\mu\beta^2}{8(L-\mu)}\left\|\dot\xi\right\|^2\mathrm{d}t + \frac{1}{8}\int_0^T \lambda_2\frac{2\beta L+\alpha}{L-\mu}\partial_t\left\|\dot\xi\right\|^2\mathrm{d}t\\ &\quad + \frac{L}{4}\int_0^T \lambda_2\frac{\alpha+2\beta\mu}{L-\mu}\partial_t\|\xi\|^2\mathrm{d}t + \frac{L\mu}{2(L-\mu)}\int_0^T \lambda_2\|\xi\|^2\mathrm{d}t\\ &\quad + \frac{L}{2(L-\mu)}\int_0^T\left(\partial_t\Big(\lambda_2\left\langle\dot\xi,\xi\right\rangle\Big) - \dot\lambda_2\left\langle\dot\xi,\xi\right\rangle - \lambda_2\left\|\dot\xi\right\|^2\right)\mathrm{d}t + \frac{1}{8(L-\mu)}\int_0^T \lambda_2\left\|\ddot\xi\right\|^2\mathrm{d}t\\ &= \int_0^T \lambda_2\phi\,\mathrm{d}t + \int_0^T \lambda_2\frac{4\alpha\beta L+\alpha^2+4L\mu\beta^2}{8(L-\mu)}\left\|\dot\xi\right\|^2\mathrm{d}t + \frac{L\mu}{2(L-\mu)}\int_0^T \lambda_2\|\xi\|^2\mathrm{d}t\\ &\quad + \frac{1}{8}\left[\lambda_2(t)\frac{2\beta(t)L+\alpha(t)}{L-\mu}\left\|\dot\xi(t)\right\|^2\right]_0^T - \frac{1}{8}\int_0^T \partial_t\left(\lambda_2\frac{2\beta L+\alpha}{L-\mu}\right)\left\|\dot\xi\right\|^2\mathrm{d}t\\ &\quad + \frac{L}{4}\left[\lambda_2(t)\frac{\alpha(t)+2\beta(t)\mu}{L-\mu}\|\xi(t)\|^2\right]_0^T - \frac{L}{4}\int_0^T \partial_t\left(\lambda_2\frac{\alpha+2\beta\mu}{L-\mu}\right)\|\xi\|^2\mathrm{d}t\\ &\quad + \frac{L}{2(L-\mu)}\left[\lambda_2(t)\left\langle\dot\xi(t),\xi(t)\right\rangle\right]_0^T - \frac{L}{4(L-\mu)}\int_0^T \dot\lambda_2\partial_t\|\xi\|^2\mathrm{d}t - \frac{L}{2(L-\mu)}\int_0^T \lambda_2\left\|\dot\xi\right\|^2\mathrm{d}t\\ &\quad + \frac{1}{8(L-\mu)}\int_0^T \lambda_2\left\|\ddot\xi\right\|^2\mathrm{d}t\\ &= \lambda_2(T)\frac{\alpha(T)+2\beta(T)L}{8(L-\mu)}\left\|\dot\xi(T)\right\|^2 + \lambda_2(T)L\frac{\alpha(T)+2\beta(T)\mu}{4(L-\mu)}\|\xi(T)\|^2 + \frac{\lambda_2(T)L}{2(L-\mu)}\left\langle\dot\xi(T),\xi(T)\right\rangle\end{aligned}$$

$$
\begin{aligned}
&- \lambda_2(0)L\frac{\alpha(0)+2\mu\beta(0)}{4(L-\mu)}\|\xi(0)\|^2 + \int_0^T \lambda_2\phi\,\mathrm{d}t + \frac{1}{8(L-\mu)}\int_0^T \lambda_2\left\|\ddot{\xi}\right\|^2\mathrm{d}t\\
&+ \frac{1}{8(L-\mu)}\int_0^T \left(\lambda_2\big(4\alpha\beta L+\alpha^2+4L\mu\beta^2\big) - \partial_t(\lambda_2(\alpha+2\beta L)) - 4L\lambda_2\right)\left\|\dot{\xi}\right\|^2\mathrm{d}t\\
&+ \frac{L}{4(L-\mu)}\int_0^T (2\mu\lambda_2 - \partial_t(\lambda_2(\alpha+2\beta\mu)))\|\xi\|^2\mathrm{d}t\\
&- \frac{L}{4(L-\mu)}\left[\dot{\lambda}_2\|\xi\|^2\right]_0^T + \frac{L}{4(L-\mu)}\int_0^T \ddot{\lambda}_2\|\xi\|^2\mathrm{d}t\\
=\ &\lambda_2(T)\frac{\alpha(T)+2\beta(T)L}{8(L-\mu)}\left\|\dot{\xi}(T)\right\|^2 + \frac{\lambda_2(T)(\alpha(T)+2\beta(T)\mu)-\dot{\lambda}_2(T)}{4(L-\mu)}L\|\xi(T)\|^2\\
&+ \frac{\lambda_2(T)L}{2(L-\mu)}\left\langle\dot{\xi}(T),\xi(T)\right\rangle - \frac{\lambda_2(0)(\alpha(0)+2\beta(0)\mu)-\dot{\lambda}_2(0)}{4(L-\mu)}L\|\xi(0)\|^2\\
&+ \int_0^T \lambda_2\phi\,\mathrm{d}t + \frac{1}{8(L-\mu)}\int_0^T \lambda_2\left\|\ddot{\xi}\right\|^2\mathrm{d}t\\
&+ \frac{1}{8(L-\mu)}\int_0^T \left(\lambda_2\big(4\alpha\beta L+\alpha^2+4L\mu\beta^2-4L\big) - \partial_t(\lambda_2(\alpha+2\beta L))\right)\left\|\dot{\xi}\right\|^2\mathrm{d}t\\
&+ \frac{L}{4(L-\mu)}\int_0^T \Big(2\mu\lambda_2 - \partial_t(\lambda_2(\alpha+2\beta\mu)) + \ddot{\lambda}_2\Big)\|\xi\|^2\mathrm{d}t.
\end{aligned}
$$

Summing up the above results, we rewrite the Lagrange function as follows:

$$
\begin{aligned}
&\mathcal{L}(\xi,\phi,\gamma(T);\lambda_1,\lambda_2,\lambda_3,\eta_1,\eta_2)\\
&\quad= \eta_1 R_1 + \eta_2 R_2 + \mathcal{P}(\xi+x^\star,\alpha,\beta)\\
&\qquad+ (\lambda_3-\lambda_1(T))\phi(T) + (\lambda_1(0)-\eta_1)\phi(0)\\
&\qquad- \frac{1}{2}\left(\frac{\lambda_1(T)}{2}\Big(1+\dot{\beta}(T)+\beta(T)\alpha(T)\Big) + \lambda_2(T)\frac{\alpha(T)+2\beta(T)L}{4(L-\mu)}\right)\left\|\dot{\xi}(T)\right\|^2\\
&\qquad- \frac{\lambda_2(T)(\alpha(T)+2\beta(T)\mu)-\dot{\lambda}_2(T)+2\lambda_3\mu}{4(L-\mu)}L\|\xi(T)\|^2\\
&\qquad- \frac{\lambda_2(T)L}{2(L-\mu)}\left\langle\dot{\xi}(T),\xi(T)\right\rangle - \frac{\lambda_3}{2(L-\mu)}\|\gamma(T)\|^2 + \frac{\lambda_3\mu}{L-\mu}\langle\gamma(T),\xi(T)\rangle\\
&\qquad+ \left(\frac{\lambda_2(0)(\alpha(0)+2\beta(0)\mu)-\dot{\lambda}_2(0)}{4(L-\mu)}L - \eta_2\right)\|\xi(0)\|^2\\
&\qquad+ \int_0^T \Big(\dot{\lambda}_1-\lambda_2\Big)\phi\,\mathrm{d}t - \frac{1}{2}\int_0^T \left(\lambda_1\beta + \frac{\lambda_2}{4(L-\mu)}\right)\left\|\ddot{\xi}\right\|^2\mathrm{d}t\\
&\qquad- \int_0^T \left(\lambda_1\frac{1+\dot{\beta}}{2}\alpha - \frac{1}{4}\partial_t\Big(\lambda_1\Big(1+\dot{\beta}+\beta\alpha\Big)\Big) + \frac{\lambda_2\big(4\alpha\beta L+\alpha^2+4L\mu\beta^2-4L\big) - \partial_t(\lambda_2(\alpha+2\beta L))}{8(L-\mu)}\right)\left\|\dot{\xi}\right\|^2\mathrm{d}t\\
&\qquad- \frac{L}{4(L-\mu)}\int_0^T \Big(2\mu\lambda_2 - \partial_t(\lambda_2(\alpha+2\beta\mu)) + \ddot{\lambda}_2\Big)\|\xi\|^2\mathrm{d}t.
\end{aligned}
$$

The Lagrange dual problem is written as

$$
\begin{aligned}
\underset{\lambda_1,\lambda_2,\lambda_3,\eta_1,\eta_2}{\text{minimize}} \quad & \max_{\phi,\xi\ \text{s.t.}\dot{\xi}(0)=0} \mathcal{L}(\xi,\phi;\lambda_1,\lambda_2,\lambda_3,\eta_1,\eta_2)\\
\text{subject to} \quad & \lambda_2(t)\geq 0, \lambda_3\geq 0, \eta_1\geq 0, \eta_2\geq 0.
\end{aligned}
$$

Since the specific modification of the dual problem above depends on the concrete form of $\mathcal{P}$, each case is discussed in the following sections. Below, for brevity, the problem obtained by modifying the Lagrange dual problem is also written simply as a dual problem.

B.2.1 Case $\mu = 0$ and $\mathcal{P}(\xi + x^\star, \alpha, \beta) = T^2\phi(T)$ (Bound on objective function values for convex functions)

In this section, we show the first inequality in Theorem 3.5. To this end, we consider the case $\mu = 0$ and $\mathcal{P}(\xi + x^\star, \alpha, \beta) = T^2\phi(T)$. Then, the Lagrange function can be written as

$$
\begin{aligned}
&\mathcal{L}(\xi, \phi, \gamma(T); \lambda_1, \lambda_2, \lambda_3, \eta_1, \eta_2) \\
&\quad = \eta_1 R_1 + \eta_2 R_2 \\
&\qquad + \big(\lambda_3 - \lambda_1(T) + T^2\big)\phi(T) + (\lambda_1(0) - \eta_1)\phi(0) \\
&\qquad - \frac{1}{2}\left(\frac{\lambda_1(T)}{2}\Big(1 + \dot\beta(T) + \beta(T)\alpha(T)\Big) + \lambda_2(T)\frac{\alpha(T) + 2\beta(T)L}{4L}\right)\left\|\dot\xi(T)\right\|^2 \\
&\qquad - \frac{\lambda_2(T)\alpha(T) - \dot\lambda_2(T)}{4}\|\xi(T)\|^2 - \frac{\lambda_2(T)}{2}\Big\langle\dot\xi(T), \xi(T)\Big\rangle - \frac{\lambda_3}{2L}\|\gamma(T)\|^2 \\
&\qquad + \left(\frac{\lambda_2(0)\alpha(0) - \dot\lambda_2(0)}{4} - \eta_2\right)\|\xi(0)\|^2 \\
&\qquad + \int_0^T \Big(\dot\lambda_1 - \lambda_2\Big)\phi\,\mathrm{d}t - \frac{1}{2}\int_0^T \left(\lambda_1\beta + \frac{\lambda_2}{4L}\right)\left\|\ddot\xi\right\|^2\mathrm{d}t \\
&\qquad - \int_0^T \left(\lambda_1\frac{1 + \dot\beta}{2}\alpha - \frac{1}{4}\partial_t\Big(\lambda_1\Big(1 + \dot\beta + \beta\alpha\Big)\Big) + \frac{\lambda_2\big(4\alpha\beta L + \alpha^2 - 4L\big) - \partial_t(\lambda_2(\alpha + 2\beta L))}{8L}\right)\left\|\dot\xi\right\|^2\mathrm{d}t \\
&\qquad - \frac{1}{4}\int_0^T \Big(-\partial_t\Big(\lambda_2\alpha - \dot\lambda_2\Big)\Big)\|\xi\|^2\mathrm{d}t.
\end{aligned}
$$

Therefore, the Lagrange dual problem reads

$$
\begin{aligned}
\underset{\lambda_1,\lambda_2,\lambda_3,\eta_1,\eta_2}{\text{minimize}} \quad & \eta_1 R_1 + \eta_2 R_2 \\
\text{subject to} \quad & \lambda_3 - \lambda_1(T) + T^2 = 0, \quad \lambda_1(0) - \eta_1 = 0, \\
& \begin{bmatrix} \lambda_1(T)\Big(1 + \dot\beta(T) + \beta(T)\alpha(T)\Big) + \lambda_2(T)\frac{\alpha(T) + 2\beta(T)L}{2L} & \lambda_2(T) \\ \lambda_2(T) & \lambda_2(T)\alpha(T) - \dot\lambda_2(T) \end{bmatrix} \succeq O, \\
& \frac{1}{4}\Big(\lambda_2(0)\alpha(0) - \dot\lambda_2(0)\Big) - \eta_2 \le 0, \\
& \dot\lambda_1(t) - \lambda_2(t) = 0, \quad \lambda_1(t)\beta(t) + \frac{\lambda_2(t)}{2L} \ge 0, \quad \partial_t\Big(\lambda_2(t)\alpha(t) - \dot\lambda_2(t)\Big) \le 0, \\
& \lambda_1\Big(1 + \dot\beta\Big)\alpha - \frac{1}{2}\partial_t\Big(\lambda_1\Big(1 + \dot\beta + \beta\alpha\Big)\Big) + \frac{\lambda_2\big(4\alpha\beta L + \alpha^2 - 4L\big) - \partial_t(\lambda_2(\alpha + 2\beta L))}{4L} \ge 0, \\
& \lambda_2(t) \ge 0, \quad \lambda_3 \ge 0, \quad \eta_1 \ge 0, \quad \eta_2 \ge 0.
\end{aligned}
$$

Below, we try to find a feasible solution that leads to a good convergence rate. This part corresponds to (iii) in Section 2.4. By using ansatz $\alpha(t) = \frac{a}{t}$, $\beta(t) = b$, $\lambda_1(t) = t^2$, $\lambda_2(t) = 2t$ with $a, b \in \mathbb{R}$, the constraints can be simplified as

$$
\begin{aligned}
& \lambda_3 = 0, \quad \eta_1 = 0, \\
& \begin{bmatrix} T^2 + (a+2)bT + \frac{a}{L} & 2T \\ 2T & 2(a-1) \end{bmatrix} \succeq O, \\
& \frac{a-1}{2} - \eta_2 \le 0, \\
& bt^2 + \frac{t}{L} \ge 0, \quad \partial_t(2a - 2) \le 0, \\
& (a-3)t + \frac{1}{2}(3a-2)b + \frac{a^2}{2Lt^2} \ge 0, \\
& \lambda_3 \ge 0, \quad \eta_1 \ge 0, \quad \eta_2 \ge 0.
\end{aligned}
$$

As a consequence, when $a \geq 3$ and $T \geq 1$, $\eta_1 = 0$ and $\eta_2 = \frac{a-1}{2}$ is a feasible solution of the dual problem. Therefore, we obtain

$$f(x(T)) - f^\star \leq \frac{a-1}{2T^2}\|x(0) - x^\star\|^2.$$

B.2.2 Case $\mu = 0$ and $\mathcal{P}(x + x^\star, \alpha, \beta) = \int_0^T \sigma(t)\|\gamma(t)\|^2 \mathrm{d}t$ (Bound on the integration of gradient norm for convex functions)

In this section, we show the second inequality in Theorem 3.5. To this end, we consider the case $\mu = 0$ and $\mathcal{P}(x + x^\star, \alpha, \beta) = \int_0^T \sigma(t)\|\gamma(t)\|^2 \mathrm{d}t$, where $\sigma\colon \mathbb{R} \to \mathbb{R}$ is a function and the specific definition of it will be given later. Then, by using the ODE (1), we have

$$\begin{aligned}
\mathcal{P}(x + x^\star, \alpha, \beta) &= \frac{1}{4}\int_0^T \sigma \left\|\ddot{\xi} + \alpha\dot{\xi}\right\|^2 \mathrm{d}t \\
&= \frac{1}{4}\int_0^T \sigma \left\|\ddot{\xi}\right\|^2 \mathrm{d}t + \frac{1}{4}\int_0^T \sigma\alpha^2 \left\|\dot{\xi}\right\|^2 \mathrm{d}t + \frac{1}{4}\int_0^T \sigma\alpha\partial_t \left\|\dot{\xi}\right\|^2 \mathrm{d}t \\
&= \frac{1}{4}\int_0^T \sigma \left\|\ddot{\xi}\right\|^2 \mathrm{d}t + \frac{1}{4}\int_0^T \sigma\alpha^2 \left\|\dot{\xi}\right\|^2 \mathrm{d}t + \frac{1}{4}\left[\sigma\alpha \left\|\dot{\xi}\right\|^2\right]_0^T - \frac{1}{4}\int_0^T \partial_t(\sigma\alpha)\left\|\dot{\xi}\right\|^2 \mathrm{d}t \\
&= \frac{1}{4}\int_0^T \sigma \left\|\ddot{\xi}\right\|^2 \mathrm{d}t + \frac{1}{4}\int_0^T \left(\sigma\alpha^2 - \partial_t(\sigma\alpha)\right)\left\|\dot{\xi}\right\|^2 \mathrm{d}t + \frac{1}{4}\sigma(T)\alpha(T)\left\|\dot{\xi}(T)\right\|^2.
\end{aligned}$$

The Lagrange function can be written as

$$\begin{aligned}
&\mathcal{L}(x, \phi, \gamma(T); \lambda_1, \lambda_2, \lambda_3, \eta_1, \eta_2) \\
&\quad = \eta_1 R_1 + \eta_2 R_2 \\
&\qquad + (\lambda_3 - \lambda_1(T))\phi(T) + (\lambda_1(0) - \eta_1)\phi(0) \\
&\qquad - \frac{1}{2}\left(\frac{\lambda_1(T)}{2}\left(1 + \dot{\beta}(T) + \beta(T)\alpha(T)\right) + \lambda_2(T)\frac{\alpha(T) + 2\beta(T)L}{4L} - \frac{1}{2}\sigma(T)\alpha(T)\right)\left\|\dot{\xi}(T)\right\|^2 \\
&\qquad - \frac{\lambda_2(T)\alpha(T) - \dot{\lambda}_2(T)}{4}\|\xi(T)\|^2 - \frac{\lambda_2(T)}{2}\left\langle \dot{\xi}(T), \xi(T)\right\rangle - \frac{\lambda_3}{2L}\|\gamma(T)\|^2 \\
&\qquad + \left(\frac{\lambda_2(0)\alpha(0) - \dot{\lambda}_2(0)}{4} - \eta_2\right)\|\xi(0)\|^2 \\
&\qquad + \int_0^T \left(\dot{\lambda}_1 - \lambda_2\right)\phi \,\mathrm{d}t - \frac{1}{2}\int_0^T \left(\lambda_1\beta + \frac{\lambda_2}{4L} - \frac{\sigma}{2}\right)\left\|\ddot{\xi}\right\|^2 \mathrm{d}t - \frac{1}{4}\int_0^T \left(-\partial_t(\lambda_2\alpha) + \ddot{\lambda}_2\right)\|\xi\|^2 \mathrm{d}t \\
&\qquad - \int_0^T \left(\lambda_1\frac{1 + \dot{\beta}}{2}\alpha - \frac{1}{4}\partial_t\left(\lambda_1\left(1 + \dot{\beta} + \beta\alpha\right)\right)\right. \\
&\qquad\qquad \left. + \frac{\lambda_2\left(4\alpha\beta L + \alpha^2 - 4L\right) - \partial_t(\lambda_2(\alpha + 2\beta L))}{8L} - \frac{1}{4}\sigma\alpha^2 + \frac{1}{4}\partial_t(\sigma\alpha)\right)\left\|\dot{\xi}\right\|^2 \mathrm{d}t.
\end{aligned}$$

Therefore, the Lagrange dual problem reads

$$\begin{aligned}
&\underset{\lambda_1, \lambda_2, \lambda_3, \eta_1, \eta_2}{\text{minimize}} \quad \eta_1 R_1 + \eta_2 R_2 \\
&\text{subject to} \\
&\lambda_3 - \lambda_1(T) = 0, \quad \lambda_1(0) - \eta_1 = 0, \\
&\begin{bmatrix} \lambda_1(T)\left(1 + \dot{\beta}(T) + \beta(T)\alpha(T)\right) + \lambda_2(T)\frac{\alpha(T) + 2\beta(T)L}{2L} - \sigma(T)\alpha(T) & \lambda_2(T) \\ \lambda_2(T) & \lambda_2(T)\alpha(T) - \dot{\lambda}_2(T) \end{bmatrix} \succeq O, \\
&\frac{1}{4}\left(\lambda_2(0)\alpha(0) - \dot{\lambda}_2(0)\right) - \eta_2 \leq 0,
\end{aligned}$$

$$\dot{\lambda}_1(t) - \lambda_2(t) = 0, \quad \lambda_1(t)\beta(t) + \frac{\lambda_2(t)}{4L} - \frac{\sigma(t)}{2} \geq 0, \quad -\partial_t(\lambda_2(t)\alpha(t)) + \ddot{\lambda}_2(t) \geq 0,$$
$$2\lambda_1\Big(1+\dot{\beta}\Big)\alpha - \partial_t\Big(\lambda_1\Big(1+\dot{\beta}+\beta\alpha\Big)\Big) + \frac{\lambda_2\big(4\alpha\beta L + \alpha^2 - 4L\big) - \partial_t(\lambda_2(\alpha+2\beta L))}{2L} - \sigma\alpha^2 + \partial_t(\sigma\alpha) \geq 0,$$
$$\lambda_2(t) \geq 0, \quad \lambda_3 \geq 0, \quad \eta_1 \geq 0, \quad \eta_2 \geq 0.$$

By eliminating λ_2, λ_3 and η_1 by $\lambda_2(t) = \dot{\lambda}_1(t)$, $\lambda_3 = \lambda_1(T)$ and $\eta_1 = \lambda_1(0)$, respectively, we have

$$\begin{aligned}
\underset{\lambda_1,\eta_2}{\text{minimize}} \quad & \lambda_1(0)R_1 + \eta_2 R_2 \\
\text{subject to} \quad & \begin{bmatrix} \lambda_1(T)\Big(1+\dot{\beta}(T)+\beta(T)\alpha(T)\Big) + \dot{\lambda}_1(T)\frac{\alpha(T)+2\beta(T)L}{2L} - \sigma(T)\alpha(T) & \dot{\lambda}_1(T) \\ \dot{\lambda}_1(T) & \dot{\lambda}_1(T)\alpha(T) - \ddot{\lambda}_1(T) \end{bmatrix} \succeq O, \\
& \frac{1}{4}\Big(\dot{\lambda}_1(0)\alpha(0) - \ddot{\lambda}_1(0)\Big) - \eta_2 \leq 0, \\
& \lambda_1(t)\beta(t) + \frac{\dot{\lambda}_1(t)}{4L} - \frac{\sigma(t)}{2} \geq 0, \quad -\partial_t\Big(\dot{\lambda}_1(t)\alpha(t) - \ddot{\lambda}_1(t)\Big) \geq 0, \\
& -\dot{\lambda}_1\left(3 + 2\dot{\beta} - \alpha\beta + \frac{\dot{\alpha}-\alpha^2}{2L}\right) + \lambda_1\Big(2\alpha - \ddot{\beta} + \alpha\dot{\beta} - \dot{\alpha}\beta\Big) - \ddot{\lambda}_1\Big(\beta + \frac{\alpha}{2L}\Big) - \sigma\alpha^2 + \partial_t(\sigma\alpha) \geq 0, \\
& \dot{\lambda}_1(t) \geq 0, \quad \lambda_1(T) \geq 0, \quad \lambda_1(0) \geq 0, \quad \eta_2 \geq 0.
\end{aligned}$$

Below, we try to find a feasible solution that leads to a good convergence rate. This part corresponds to (iii) in Section 2.4. By using ansatz $\alpha(t) = \frac{a}{t}$, $\beta(t) = b$ and $\lambda_1(t) = ct^2$ with $a, b \in \mathbb{R}$ and $c \geq 0$, the constraints can be simplified as

$$\begin{aligned}
& \begin{bmatrix} cT^2 + (a+2)bcT + \frac{ac}{L} - \frac{a\sigma(T)}{T} & 2cT \\ 2cT & 2c(a-1) \end{bmatrix} \succeq O, \\
& 2bct^2 + \frac{ct}{L} - \sigma(t) \geq 0, \\
& 2(a-3)ct + (3a-2)bc + \frac{a^2c}{L}\frac{1}{t} - \frac{a(a+1)}{t^2}\sigma(t) + \frac{a}{t}\dot{\sigma}(t) \geq 0.
\end{aligned}$$

By choosing $\sigma(t) = c\Big(\frac{3a-2}{a(a-1)}bt^2 + \frac{t}{L}\Big)$, the constraints can be further simplified as

$$\begin{bmatrix} T^2 + \frac{a(a-2)}{a-1}bT & 2T \\ 2T & 2(a-1) \end{bmatrix} \succeq O, \qquad \frac{(a-2)(2a-1)}{a(a-1)}bt^2 \geq 0, \qquad 2(a-3)t \geq 0.$$

Since the first condition holds if and only if $T^2 + \frac{a(a-2)}{a-1}bT \geq 0$, $2(a-1) \geq 0$ and $2(a-1)T^2 + a(a-2)bT - 4T^2 = 2(a-3)T^2 + a(a-2)bT \geq 0$ hold, for any $\alpha(t) = \frac{a}{t}$ and $\beta(t) = b$ with $a \geq 3$ and $b \geq 0$, respectively, $\lambda_1(t) = t^2$, $\eta_2 = \frac{a-1}{2}$ is a feasible solution of the dual problem. Therefore, we obtain

$$\int_0^T \left(\frac{3a-2}{a(a-1)^2}bt^2 + \frac{t}{(a-1)L}\right)\|\nabla f(x(t) + b\dot{x})\|^2 \mathrm{d}t \leq \frac{1}{2}\|x_0 - x^\star\|^2.$$

The factor $\frac{3a-2}{a(a-1)^2}$ is maximized at $a = 3$ (recall that $a \geq 3$), and then it is equal to $\frac{7}{12}$.

B.2.3 Case $\mu = 0$ and $\mathcal{P}(\xi + x^\star, \alpha, \beta) = \int_0^T \dot{\sigma}(t)\phi(t)\mathrm{d}t$ (Bound on the integration of objective function values for convex functions)

In this section, we show the last inequality in Theorem 3.5. To this end, we consider the case $\mu = 0$ and $\mathcal{P}(\xi + x^\star, \alpha, \beta) = \int_0^T \dot{\sigma}(t)\phi(t)\mathrm{d}t$, where $\sigma\colon \mathbb{R} \to \mathbb{R}$ is a function and the specific definition of it will be given later. Then, the Lagrange function can be written as

$$\mathcal{L}(\xi, \phi, \gamma(T); \lambda_1, \lambda_2, \lambda_3, \eta_1, \eta_2)$$

$$
\begin{aligned}
&= \eta_1 R_1 + \eta_2 R_2 \\
&\quad + (\lambda_3 - \lambda_1(T))\phi(T) + (\lambda_1(0) - \eta_1)\phi(0) \\
&\quad - \frac{1}{2}\left(\frac{\lambda_1(T)}{2}\Big(1 + \dot\beta(T) + \beta(T)\alpha(T)\Big) + \lambda_2(T)\frac{\alpha(T) + 2\beta(T)L}{4L}\right)\Big\|\dot\xi(T)\Big\|^2 \\
&\quad - \frac{\lambda_2(T)\alpha(T) - \dot\lambda_2(T)}{4}\|\xi(T)\|^2 - \frac{\lambda_2(T)}{2}\Big\langle \dot\xi(T), \xi(T)\Big\rangle - \frac{\lambda_3}{2L}\|\gamma(T)\|^2 \\
&\quad + \left(\frac{\lambda_2(0)\alpha(0) - \dot\lambda_2(0)}{4} - \eta_2\right)\|\xi(0)\|^2 \\
&\quad + \int_0^T \Big(\dot\lambda_1 - \lambda_2 + \dot\sigma\Big)\phi \,\mathrm{d}t - \frac{1}{2}\int_0^T \left(\lambda_1\beta + \frac{\lambda_2}{4L}\right)\Big\|\ddot\xi\Big\|^2 \mathrm{d}t \\
&\quad - \int_0^T \left(\lambda_1\frac{1 + \dot\beta}{2}\alpha - \frac{1}{4}\partial_t\Big(\lambda_1\Big(1 + \dot\beta + \beta\alpha\Big)\Big) + \frac{\lambda_2\big(4\alpha\beta L + \alpha^2 - 4L\big) - \partial_t(\lambda_2(\alpha + 2\beta L))}{8L}\right)\Big\|\dot\xi\Big\|^2 \mathrm{d}t \\
&\quad - \frac{1}{4}\int_0^T \Big(-\partial_t\Big(\lambda_2\alpha - \dot\lambda_2\Big)\Big)\|\xi\|^2 \mathrm{d}t.
\end{aligned}
$$

Therefore, the Lagrange dual problem reads

$$
\begin{aligned}
\underset{\lambda_1,\lambda_2,\lambda_3,\eta_1,\eta_2}{\text{minimize}} \quad & \eta_1 R_1 + \eta_2 R_2 \\
\text{subject to} \quad & \lambda_3 - \lambda_1(T) = 0, \quad \lambda_1(0) - \eta_1 = 0, \\
& \begin{bmatrix} \lambda_1(T)\Big(1 + \dot\beta(T) + \beta(T)\alpha(T)\Big) + \lambda_2(T)\frac{\alpha(T) + 2\beta(T)L}{2L} & \lambda_2(T) \\ \lambda_2(T) & \lambda_2(T)\alpha(T) - \dot\lambda_2(T) \end{bmatrix} \succeq O, \\
& \frac{1}{4}\Big(\lambda_2(0)\alpha(0) - \dot\lambda_2(0)\Big) - \eta_2 \le 0, \\
& \dot\lambda_1(t) - \lambda_2(t) + \dot\sigma(t) = 0, \quad \lambda_1(t)\beta(t) + \frac{\lambda_2(t)}{2L} \ge 0, \quad \partial_t\Big(\lambda_2(t)\alpha(t) - \dot\lambda_2(t)\Big) \le 0, \\
& \lambda_1\Big(1 + \dot\beta\Big)\alpha - \frac{1}{2}\partial_t\Big(\lambda_1\Big(1 + \dot\beta + \beta\alpha\Big)\Big) + \frac{\lambda_2\big(4\alpha\beta L + \alpha^2 - 4L\big) - \partial_t(\lambda_2(\alpha + 2\beta L))}{4L} \ge 0, \\
& \lambda_2(t) \ge 0, \quad \lambda_3 \ge 0, \quad \eta_1 \ge 0, \quad \eta_2 \ge 0.
\end{aligned}
$$

Below, we try to find a feasible solution that leads to a good convergence rate. This part corresponds to (iii) in Section 2.4. By using ansatz $\alpha(t) = \frac{a}{t}$, $\beta(t) = b$, $\lambda_1(t) = (1-c)t^2$, $\lambda_2(t) = 2t$ and $\sigma(t) = ct^2$ with $a, b \in \mathbb{R}$ and $c \in [0,1]$, the constraints can be simplified as

$$
\begin{aligned}
& \lambda_3 = (1-c)T^2, \quad \eta_1 = 0, \\
& \begin{bmatrix} (1-c)\big(T^2 + abT\big) + \frac{a}{L} + 2bT & 2T \\ 2T & 2(a-1) \end{bmatrix} \succeq O, \\
& \frac{a-1}{2} - \eta_2 \le 0, \quad b(1-c)t^2 + \frac{t}{L} \ge 0, \\
& (a - 3 - c(a-1))t + \frac{ac + 3a - 8}{2}b + \frac{a^2}{2tL} \ge 0, \\
& 2t \ge 0, \quad \lambda_3 \ge 0, \quad \eta_2 \ge 0.
\end{aligned}
$$

As a consequence, for any $a > 3$, $c = \frac{a-3}{a-1}$ gives a feasible solution of the dual problem. Therefore, we obtain

$$
\int_0^T 2\frac{a-3}{a-1}t(f(x(t)) - f^\star)\mathrm{d}t \le \frac{a-1}{2},
$$

which implies

$$
\int_0^T t(f(x(t)) - f^\star)\mathrm{d}t \le \frac{(a-1)^2}{4(a-3)}.
$$

The right-hand side is minimized at $a = 5$, and then it is equal to 2.

B.2.4 Case $\mu > 0$, $L = \infty$ and $\mathcal{P}(\xi + x^\star, \alpha, \beta) = \phi(T)$ (Bound on objective function values for strongly convex functions)

In this section, we prove the second halves in Theorems 3.1 and 3.3. To this end, we consider the case $\mu > 0$, $L = \infty$, and $\mathcal{P}(\xi + x^\star, \alpha, \beta) = \phi(T)$. Then, the Lagrange function can be written as

$$
\begin{aligned}
&\mathcal{L}(\xi, \phi, \gamma(T); \lambda_1, \lambda_2, \lambda_3, \eta_1, \eta_2) \\
&\quad = \eta_1 R_1 + \eta_2 R_2 \\
&\qquad + (\lambda_3 - \lambda_1(T) + 1)\phi(T) + (\lambda_1(0) - \eta_1)\phi(0) \\
&\qquad - \frac{1}{2}\left(\frac{\lambda_1(T)}{2}\Big(1 + \dot\beta(T) + \beta(T)\alpha(T)\Big) + \frac{\lambda_2(T)\beta(T)}{2}\right)\Big\|\dot\xi(T)\Big\|^2 \\
&\qquad - \frac{\lambda_2(T)(\alpha(T) + 2\beta(T)\mu) - \dot\lambda_2(T) + 2\lambda_3\mu}{4}\|\xi(T)\|^2 - \frac{\lambda_2(T)}{2}\Big\langle \dot\xi(T), \xi(T)\Big\rangle \\
&\qquad + \left(\frac{\lambda_2(0)(\alpha(0) + 2\beta(0)\mu) - \dot\lambda_2(0)}{4} - \eta_2\right)\|\xi(0)\|^2 \\
&\qquad + \int_0^T \Big(\dot\lambda_1 - \lambda_2\Big)\phi \,\mathrm{d}t - \frac{1}{2}\int_0^T (\lambda_1\beta)\Big\|\ddot\xi\Big\|^2 \mathrm{d}t \\
&\qquad - \int_0^T \left(\lambda_1 \frac{1 + \dot\beta}{2}\alpha - \frac{1}{4}\partial_t\Big(\lambda_1\Big(1 + \dot\beta + \beta\alpha\Big)\Big) + \frac{2\lambda_2\big(\alpha\beta + \mu\beta^2 - 1\big) - \partial_t(\lambda_2\beta)}{4}\right)\Big\|\dot\xi\Big\|^2 \mathrm{d}t \\
&\qquad - \frac{1}{4}\int_0^T \Big(2\mu\lambda_2 - \partial_t(\lambda_2(\alpha + 2\beta\mu)) + \ddot\lambda_2\Big)\|\xi\|^2 \mathrm{d}t.
\end{aligned}
$$

The Lagrange dual problem reads

$$
\begin{aligned}
\underset{\lambda_1,\lambda_2,\lambda_3,\eta_1,\eta_2}{\text{minimize}} \quad & \eta_1 R_1 + \eta_2 R_2 \qquad (30)\\
\text{subject to} \quad & \lambda_3 - \lambda_1(T) + 1 = 0, \quad \lambda_1(0) - \eta_1 = 0, \\
& \begin{bmatrix} \lambda_1(T)\Big(1 + \dot\beta(T) + \beta(T)\alpha(T)\Big) + \lambda_2(T)\beta(T) & \lambda_2(T) \\ \lambda_2(T) & \lambda_2(T)(\alpha(T) + 2\beta(T)\mu) - \dot\lambda_2(T) + 2\lambda_3\mu \end{bmatrix} \succeq O, \\
& \frac{\lambda_2(0)(\alpha(0) + 2\beta(0)\mu) - \dot\lambda_2(0)}{4} - \eta_2 \le 0, \\
& \dot\lambda_1(t) - \lambda_2(t) = 0, \quad \lambda_1(t)\beta(t) \ge 0, \quad 2\mu\lambda_2(t) - \partial_t(\lambda_2(t)(\alpha(t) + 2\beta(t)\mu)) + \ddot\lambda_2(t) \ge 0, \\
& \lambda_1\Big(1 + \dot\beta\Big)\alpha - \frac{1}{2}\partial_t\Big(\lambda_1\Big(1 + \dot\beta + \beta\alpha\Big)\Big) + \frac{2\lambda_2\big(\alpha\beta + \mu\beta^2 - 1\big) - \partial_t(\lambda_2\beta)}{2} \ge 0, \\
& \lambda_2(t) \ge 0, \quad \lambda_3 \ge 0, \quad \eta_1 \ge 0, \quad \eta_2 \ge 0.
\end{aligned}
$$

Strongly convex functions: exponential rate We try to find a feasible solution that leads to a good convergence rate. This part corresponds to (iii) in Section 2.4. By using ansatz $\alpha(t) = \sigma + \frac{2\mu}{\sigma}$, $\beta(t) = 0$, $\lambda_1(t) = (1 + \lambda_3)e^{\sigma(t-T)}$ and $\lambda_2(t) = (1 + \lambda_3)\sigma e^{\sigma(t-T)}$ with $\sigma \ge 0$, the constraints can be simplified as

$$
\begin{aligned}
& (1 + \lambda_3)e^{-\sigma T} - \eta_1 = 0, \\
& \begin{bmatrix} 1 + \lambda_3 & \sigma(1 + \lambda_3) \\ \sigma(1 + \lambda_3) & 2\mu(1 + 2\lambda_3) \end{bmatrix} \succeq O, \\
& \frac{(1 + \lambda_3)\mu}{2}e^{-\sigma T} - \eta_2 \le 0, \\
& (1 + \lambda_3)\frac{1}{2\sigma}\big(4\mu - \sigma^2\big)e^{\sigma(t-T)} \ge 0, \\
& \lambda_3 \ge 0, \quad \eta_1 \ge 0, \quad \eta_2 \ge 0.
\end{aligned}
$$

The second constraint is equivalent to $\lambda_3 + 1 \geq 0$, $2\mu(2\lambda_3+1) \geq 0$ and $2\mu(\lambda_3+1)(2\lambda_3+1) - \sigma^2(\lambda_3+1)^2 = (\lambda_3+1)\big((4\mu-\sigma^2)\lambda_3 + 2\mu - \sigma^2\big) \geq 0$. As a consequence, when $\sigma \in [\sqrt{2\mu}, 2\sqrt{\mu})$, $\lambda_1(t) = \frac{2\mu}{4\mu-\sigma^2}e^{\sigma(t-T)}$, $\lambda_2(t) = \frac{2\mu}{4\mu-\sigma^2}\sigma e^{\sigma(t-T)}$, $\lambda_3 = \frac{\sigma^2-2\mu}{4\mu-\sigma^2}$, $\eta_1 = \frac{2\mu}{4\mu-\sigma^2}e^{-\sigma T}$ and $\eta_2 = \frac{\mu^2}{4\mu-\sigma^2}e^{-\sigma T}$ is feasible solution of the dual problem. Therefore, we obtain

$$f(x(T)) - f^\star \leq \frac{2\mu}{4\mu-\sigma^2}e^{-\sigma T}\Big(f(x(0)) - f^\star + \frac{\mu}{2}\|x(0)-x^\star\|^2\Big).$$

Furthermore, since

$$\begin{bmatrix} 1+\lambda_3 & \sigma(1+\lambda_3) \\ \sigma(1+\lambda_3) & 2\mu(1+2\lambda_3) \end{bmatrix} = \frac{2\mu}{4\mu-\sigma^2}\begin{bmatrix} 1 & \sigma \\ \sigma & \sigma^2 \end{bmatrix}$$

holds for the chosen feasible solution, the same argument holds true even if we replace the objective function with

$$\phi(T) + \frac{1}{2}\frac{\mu}{4\mu-\sigma^2}\Big\|\dot\xi(T) + \sigma\xi(T)\Big\|^2.$$

(This part corresponds to (iii) in Section 2.4.) This implies

$$f(x(T)) - f^\star + \frac{\mu}{2(4\mu-\sigma^2)}\|\sigma(x(T)-x^\star) + \dot x(T)\|^2 \leq \frac{2\mu}{4\mu-\sigma^2}e^{-\sigma T}\Big(f(x(0)) - f^\star + \frac{\mu}{2}\|x(0)-x^\star\|^2\Big).$$

Both convex and strongly convex functions: hyperbolic sine rate We try to find a feasible solution that leads to a good convergence rate. This part corresponds to (iii) in Section 2.4. By using ansatz $\alpha(t) = a\coth(\sigma t)$, $\beta(t) = 0$, $\lambda_1(t) = c\sinh^2(\sigma t)$ and $\lambda_2(t) = 2c\sigma\sinh(\sigma t)\cosh(\sigma t)$ with $\sigma \geq 0$, the constraints can be simplified as

$$\begin{aligned}
&\lambda_3 - c\sinh^2(\sigma T) + 1 = 0, \quad -\eta_1 = 0,\\
&\begin{bmatrix} c\sinh^2(\sigma T) & 2c\sigma\sinh(\sigma T)\cosh(\sigma T) \\ 2c\sigma\sinh(\sigma T)\cosh(\sigma T) & 2ac\sigma\cosh^2(\sigma T) - 2c\sigma^2\big(\cosh^2(\sigma T)+\sinh^2(\sigma T)\big) + 2\lambda_3\mu \end{bmatrix} \succeq O,\\
&\frac{c\sigma(a-\sigma)}{2} - \eta_2 \leq 0,\\
&4c\sigma\big(\mu - a\sigma + 2\sigma^2\big)\sinh(\sigma t)\cosh(\sigma t) \geq 0,\\
&c(a-3\sigma)\sinh(\sigma t)\cosh(\sigma t) \geq 0,\\
&2c\sigma\sinh(\sigma t)\cosh(\sigma t) \geq 0, \quad \lambda_3 \geq 0, \quad \eta_1 \geq 0, \quad \eta_2 \geq 0.
\end{aligned}$$

Since we have $c = (\lambda_3+1)\sinh^{-2}(\sigma T)$ from the first constraint, these constraints can be further simplified as

$$\begin{aligned}
&\eta_1 = 0,\\
&\begin{bmatrix} \lambda_3+1 & 2\sigma(\lambda_3+1)\coth(\sigma T) \\ 2\sigma(\lambda_3+1)\coth(\sigma T) & 2(\lambda_3+1)\sigma(a-\sigma)\coth^2(\sigma T) - 2\sigma^2(\lambda_3+1) + 2\lambda_3\mu \end{bmatrix} \succeq O,\\
&\frac{(\lambda_3+1)\sigma(a-\sigma)}{2\sinh^2(\sigma T)} - \eta_2 \leq 0,\\
&\mu - \sigma(a-2\sigma) \geq 0,\\
&a - 3\sigma \geq 0,\\
&\lambda_3 \geq 0, \quad \eta_2 \geq 0.
\end{aligned}$$

The second constraint holds if the determinant of the matrix in the left-hand side is nonnegative, i.e.,

$$2(\lambda_3+1)\big((\lambda_3+1)\sigma(a-3\sigma)\coth^2(\sigma T) + \big(\mu-\sigma^2\big)\lambda_3 - \sigma^2\big) \geq 0.$$

Based on this observation, we assume $a < 3\sqrt{\mu}$ and choose $\sigma = \frac{a}{3}$ and $\lambda_3 = \frac{\sigma^2}{\mu-\sigma^2}$ as a feasible solution. Then, the value of the objective function is

$$\frac{\mu\sigma^2}{(\mu-\sigma^2)\sinh^2(\sigma T)}R_2.$$

Therefore, we obtain

$$f(x(T)) - f^\star \leq \frac{\mu\sigma^2}{(\mu-\sigma^2)\sinh^2(\sigma T)}\|x(0)-x^\star\|^2.$$

Furthermore, since

$$\begin{bmatrix} \lambda_3+1 & 2\sigma(\lambda_3+1)\coth(\sigma T) \\ 2\sigma(\lambda_3+1)\coth(\sigma T) & 2(\lambda_3+1)\sigma(a-\sigma)\coth^2(\sigma T) - 2\sigma^2(\lambda_3+1) + 2\lambda_3\mu \end{bmatrix} = \frac{\mu}{\mu-\sigma^2}\begin{bmatrix} 1 & 2\sigma\coth(\sigma T) \\ 2\sigma\coth(\sigma T) & 4\sigma^2\coth^2(\sigma T)\end{bmatrix}$$

holds for the feasible solution, the same argument holds true even if we replace the objective function with

$$\phi(T) + \frac{\mu}{4(\mu-\sigma^2)}\left\|\dot\xi(T) + 2\sigma\coth(\sigma T)\xi(T)\right\|^2.$$

(This part corresponds to (iii) in Section 2.4.) This implies

$$f(x(T)) - f^\star + \frac{\mu}{4(\mu-\sigma^2)}\|2\sigma\coth(\sigma T)(x(T)-x^\star) + \dot x(T)\|^2 \leq \frac{\mu\sigma^2}{(\mu-\sigma^2)\sinh^2(\sigma T)}\|x(0)-x^\star\|^2$$

B.2.5 Case $\mu > 0$, $L = \infty$ and $\mathcal{P}(\xi + x^\star, \alpha, \beta) = \phi(T) - \frac{\mu}{2}\|x(T)-x^\star\|^2$ (Bound on modified objective function values for strongly convex functions)

In this section, we prove the first halves in Theorems 3.1 and 3.3. To deal with the case $\sigma = 2\sqrt{\mu}$ for the strongly convex case and $\sigma = \sqrt{\mu}$ for the both convex and strongly convex case, we modify the objective function as $\mathcal{P}(\xi + x^\star, \alpha, \beta) = \phi(T) - \frac{\mu}{2}\|x(T)-x^\star\|^2$. Then, under the assumption $\beta(t) = 0$, the Lagrange function can be written as

$$\begin{aligned}
&\mathcal{L}(\xi, \phi, \gamma(T); \lambda_1, \lambda_2, \lambda_3, \eta_1, \eta_2) \\
&\quad = \eta_1 R_1 + \eta_2 R_2 + (\lambda_3 - \lambda_1(T) + 1)\phi(T) + (\lambda_1(0) - \eta_1)\phi(0) \\
&\qquad - \frac{\lambda_1(T)}{4}\left\|\dot\xi(T)\right\|^2 - \frac{\lambda_2(T)\alpha(T) - \dot\lambda_2(T) + 2\lambda_3\mu + 2\mu}{4}\|\xi(T)\|^2 - \frac{\lambda_2(T)}{2}\left\langle \dot\xi(T), \xi(T)\right\rangle \\
&\qquad + \left(\frac{\lambda_2(0)\alpha(0) - \dot\lambda_2(0)}{4} - \eta_2\right)\|\xi(0)\|^2 \\
&\qquad + \int_0^T \left(\dot\lambda_1 - \lambda_2\right)\phi\,\mathrm{d}t - \frac{1}{4}\int_0^T \left(4\lambda_1\alpha - \dot\lambda_1 - 2\lambda_2\right)\left\|\dot\xi\right\|^2\mathrm{d}t - \frac{1}{4}\int_0^T \left(2\mu\lambda_2 - \partial_t(\lambda_2\alpha) + \ddot\lambda_2\right)\|\xi\|^2\mathrm{d}t.
\end{aligned}$$

The Lagrange dual problem reads

$$\begin{aligned}
\underset{\lambda_1,\lambda_2,\lambda_3,\eta_1,\eta_2}{\text{minimize}} \quad & \eta_1 R_1 + \eta_2 R_2 \\
\text{subject to} \quad & \lambda_3 - \lambda_1(T) + 1 = 0, \quad \lambda_1(0) - \eta_1 = 0, \\
& \begin{bmatrix} \lambda_1(T) & \lambda_2(T) \\ \lambda_2(T) & \lambda_2(T)\alpha(T) - \dot\lambda_2(T) + 2\lambda_3\mu + 2\mu \end{bmatrix} \succeq O, \\
& \frac{\lambda_2(0)\alpha(0) - \dot\lambda_2(0)}{4} - \eta_2 \leq 0, \\
& \dot\lambda_1(t) - \lambda_2(t) = 0, \quad 4\lambda_1(t)\alpha(t) - \dot\lambda_1(t) - 2\lambda_2(t) \geq 0, \quad 2\mu\lambda_2(t) - \partial_t(\lambda_2(t)\alpha(t)) + \ddot\lambda_2(t) \geq 0, \\
& \lambda_2(t) \geq 0, \quad \lambda_3 \geq 0, \quad \eta_1 \geq 0, \quad \eta_2 \geq 0.
\end{aligned}$$

Strongly convex functions: exponential rate We try to find a feasible solution that leads to a good convergence rate. This part corresponds to (iii) in Section 2.4. By using ansatz $\alpha(t) = \sigma + \frac{2\mu}{\sigma}$, $\beta(t) = 0$, $\lambda_1(t) = (1+\lambda_3)e^{\sigma(t-T)}$ and $\lambda_2(t) = (1+\lambda_3)\sigma e^{\sigma(t-T)}$ with $\sigma \geq 0$, the constraints can be simplified as

$$\begin{aligned}
&(1+\lambda_3)e^{-\sigma T} - \eta_1 = 0, \\
&(\lambda_3+1)\begin{bmatrix} 1 & \sigma \\ \sigma & 4\mu \end{bmatrix} \succeq O,
\end{aligned}$$

$$\frac{(1+\lambda_3)\mu}{2}e^{-\sigma T} - \eta_2 \leq 0,$$
$$\lambda_3 \geq 0, \quad \eta_1 \geq 0, \quad \eta_2 \geq 0.$$

Since the second constraint is equivalent to $\sigma \leq 2\sqrt{\mu}$, $\lambda_3 = 0$, $\eta_1 = e^{-\sigma T}$ and $\eta_2 = \frac{\mu}{2}e^{-\sigma T}$ is a feasible solution. Therefore, we obtain

$$f(x(T)) - f^\star - \frac{\mu}{2}\|x(T) - x^\star\|^2 \leq e^{-\sigma T}\Big(f(x(0)) - f^\star + \frac{\mu}{2}\|x(0) - x^\star\|^2\Big).$$

Furthermore, since

$$(\lambda_3 + 1)\begin{bmatrix} 1 & \sigma \\ \sigma & 4\mu \end{bmatrix} \succeq \begin{bmatrix} 1 & \sigma \\ \sigma & \sigma^2 \end{bmatrix}$$

holds for the chosen feasible solution, the same argument holds true even if we replace the objective function with

$$\phi(T) - \frac{\mu}{2}\|\xi(T)\|^2 + \frac{1}{4}\Big\|\dot{\xi}(T) + \sigma\xi(T)\Big\|^2.$$

(This part corresponds to (iii) in Section 2.4.) This implies

$$f(x(T)) - f^\star - \frac{\mu}{2}\|x(T) - x^\star\|^2 + \frac{1}{4}\|\sigma(x(T) - x^\star) + \dot{x}(T)\|^2 \leq e^{-\sigma T}\Big(f(x(0)) - f^\star + \frac{\mu}{2}\|x(0) - x^\star\|^2\Big).$$

Both convex and strongly convex functions: hyperbolic sine rate We try to find a feasible solution that leads to a good convergence rate. This part corresponds to (iii) in Section 2.4. By using ansatz $\alpha(t) = a\coth(\sigma t)$, $\beta(t) = 0$, $\lambda_1(t) = c\sinh^2(\sigma t)$ and $\lambda_2(t) = 2c\sigma\sinh(\sigma t)\cosh(\sigma t)$ with $\sigma \geq 0$, the constraints can be simplified as

$$\lambda_3 - c\sinh^2(\sigma T) + 1 = 0, \quad -\eta_1 = 0,$$
$$\begin{bmatrix} c\sinh^2(\sigma T) & 2c\sigma\sinh(\sigma T)\cosh(\sigma T) \\ 2c\sigma\sinh(\sigma T)\cosh(\sigma T) & 2ac\sigma\cosh^2(\sigma T) - 2c\sigma^2\big(\cosh^2(\sigma T) + \sinh^2(\sigma T)\big) + 2\mu(\lambda_3+1) \end{bmatrix} \succeq O,$$
$$\frac{c\sigma(a-\sigma)}{2} - \eta_2 \leq 0,$$
$$4c\sigma\big(\mu - a\sigma + 2\sigma^2\big)\sinh(\sigma t)\cosh(\sigma t) \geq 0,$$
$$c(a - 3\sigma)\sinh(\sigma t)\cosh(\sigma t) \geq 0,$$
$$2c\sigma\sinh(\sigma t)\cosh(\sigma t) \geq 0, \quad \lambda_3 \geq 0, \quad \eta_1 \geq 0, \quad \eta_2 \geq 0.$$

Since we have $c = (\lambda_3 + 1)\sinh^{-2}(\sigma T)$ and $\eta_1 = 0$ from the first and second constraints, respectively, these constraints can be further simplified as

$$(\lambda_3 + 1)\begin{bmatrix} 1 & 2\sigma\coth(\sigma T) \\ 2\sigma\coth(\sigma T) & 2\sigma(a-\sigma)\coth^2(\sigma T) - 2\sigma^2 + 2\mu \end{bmatrix} \succeq O,$$
$$\frac{(\lambda_3+1)\sigma(a-\sigma)}{2\sinh(\sigma T)} - \eta_2 \leq 0, \quad \mu - \sigma(a - 2\sigma) \geq 0, \quad a - 3\sigma \geq 0.$$

The first constraint holds if the determinant of the matrix in the left-hand side is nonnegative, i.e.,

$$2\sigma(a - 3\sigma)\coth^2(\sigma T) - 2\sigma^2 + 2\mu \geq 0.$$

Baed on this observation, we assume $a \leq 3\sqrt{\mu}$ and choose $\sigma = \frac{a}{3}$ and $\lambda_3 = 0$ as a feasible solution. Then, the value of the objective function is $\frac{\sigma^2}{\sinh^2(\sigma T)}R_2$. Therefore, we obtain

$$f(x(T)) - f^\star - \frac{\mu}{2}\|x(T) - x^\star\|^2 \leq \frac{\sigma^2}{\sinh^2(\sigma T)}\|x(0) - x^\star\|^2.$$

Furthermore, since

$$(\lambda_3 + 1)\begin{bmatrix} 1 & 2\sigma\coth(\sigma T) \\ 2\sigma\coth(\sigma T) & 2\sigma(a-\sigma)\coth^2(\sigma T) - 2\sigma^2 + 2\mu \end{bmatrix} \succeq \begin{bmatrix} 1 & 2\sigma\coth(\sigma T) \\ 2\sigma\coth(\sigma T) & 4\sigma^2\coth^2(\sigma T) \end{bmatrix}$$

holds for the chosen feasible solution, the same argument holds true even if we replace the objective function with

$$\phi(T) - \frac{\mu}{2}\|\xi(T)\|^2 + \frac{1}{4}\left\|\dot{\xi}(T) + 2\sigma\coth(\sigma T)\xi(T)\right\|^2.$$

(This part corresponds to (iii) in Section 2.4.) This implies

$$f(x(T)) - f^\star - \frac{\mu}{2}\|x(T)-x^\star\|^2 + \frac{1}{4}\|2\sigma\coth(\sigma T)(x(T)-x^\star) + \dot{x}(T)\|^2 \leq \frac{\sigma^2}{\sinh^2(\sigma T)}\|x(0)-x^\star\|^2.$$

B.2.6 Case $\mu > 0$ and $L < \infty$

In this section, we demonstrate how convergence rates deteriorate by using the information $L < \infty$. We consider the TMM ODE: $\alpha(t) = 3\sqrt{\mu}$ and $\beta(t) = 0$. Similarly to Section 2, we suppose $\mathcal{P}(\xi + x^\star, \alpha, \beta)$ is a quadratic form of $\dot{\xi}(T)$, $\xi(T)$ and $\gamma(T)$. Then, the dual problem (30) reads

$$\begin{aligned}
\underset{\lambda_1,\lambda_2,\lambda_3,\eta_1,\eta_2}{\text{minimize}} \quad & \eta_1 R_1 + \eta_2 R_2 && (31)\\
\text{subject to} \quad & \lambda_3 - \lambda_1(T) = 0, \quad \lambda_1(0) - \eta_1 = 0,\\
& \begin{bmatrix} \lambda_1(T) + \lambda_2(T)\frac{\alpha(T)}{2(L-\mu)} & \frac{\lambda_2(T)L}{L-\mu} & 0 \\ \frac{\lambda_2(T)L}{L-\mu} & \frac{\lambda_2(T)\alpha(T) - \dot{\lambda}_2(T) + 2\lambda_3\mu}{L-\mu}L & \frac{2\lambda_3\mu}{L-\mu} \\ 0 & \frac{2\lambda_3\mu}{L-\mu} & \frac{2\lambda_3}{L-\mu} \end{bmatrix} - (\text{matrix associated to } \mathcal{P}) \succeq O,\\
& \frac{\lambda_2(0)\alpha(0) - \dot{\lambda}_2(0)}{4} - \eta_2 \leq 0,\\
& \dot{\lambda}_1(t) - \lambda_2(t) = 0,\\
& \lambda_1\frac{\alpha}{2} - \frac{1}{4}\dot{\lambda}_1 + \frac{\lambda_2\big(\alpha^2 - 4L\big) - \partial_t(\lambda_2\alpha)}{8(L-\mu)} \geq 0, && (32)\\
& 2\mu\lambda_2(t) - \partial_t(\lambda_2(t)\alpha(t)) + \ddot{\lambda}_2(t) \geq 0, && (33)\\
& \lambda_2(t) \geq 0, \quad \lambda_3 \geq 0, \quad \eta_1 \geq 0, \quad \eta_2 \geq 0.
\end{aligned}$$

Using ansatz $\lambda_1 = \lambda_3 \mathrm{e}^{\sigma(t-T)}$ and $\lambda_2 = \lambda_3\sigma\mathrm{e}^{\sigma(t-T)}$, the constraints (32) can be simplified as

$$\frac{3\sqrt{\mu}}{2} - \frac{3\sqrt{\mu}}{4}\sigma + \frac{(9\mu - 4L)\sigma - 3\sqrt{\mu}\sigma^2}{8(L-\mu)} \geq 0.$$

From this inequality, we have

$$\sigma \leq \frac{\sqrt{36L^2 + 12\mu L - 23\mu^2} + 11\mu - 6L}{6\sqrt{\mu}},$$

where the RHS is monotonically increasing in L/μ and equal to $\frac{5\sqrt{\mu}}{3}$ when $L = \mu$ and $2\sqrt{\mu}$ when $L = \infty$. However, from the constraint (33), we have $\sigma \leq \sqrt{\mu}$ or $2\sqrt{\mu} \leq \sigma$. Therefore, by using the information $L < \infty$, we obtain $\sigma \leq \sqrt{\mu}$ that is worse than the case we do not assume that.

Adopting another $\alpha(t)$, we obtain a mild deterioration.

Theorem B.1. *Let $f \in \mathcal{F}_{\mu,L}$ and let $x : [0,T] \to \mathbb{R}^d$ be the solution of eq. (1) with $\alpha(t) \coloneqq \sqrt{2\mu}(\sqrt{2-\mu/L} + 1/\sqrt{2-\mu/L})$ and $\beta(t) \coloneqq 0$. Then, we have for any $T \geq 0$*

$$\frac{1}{4}\|\sqrt{2\mu(2-\mu/L)}(x(T)-x^\star) + \dot{x}(T)\|^2 \leq \mathrm{e}^{-\sqrt{2\mu(2-\mu/L)}T}\left(f(x(0)) - f^\star + \frac{\mu}{2}\|x(0)-x^\star\|^2\right).$$

Proof. We set $\mathcal{P}(\xi + x^\star, \alpha, \beta) = \frac{L}{4(L-\mu)}\left\|\dot{\xi}(T) + \sigma\xi(T)\right\|^2$, and using ansatz $\alpha(t) = \sigma + \frac{2\mu}{\sigma}$, $\lambda_1 = \lambda_3\mathrm{e}^{\sigma(t-T)}$, $\lambda_2 = (1+\lambda_3)\sigma\sqrt{\mu}\mathrm{e}^{\sigma\sqrt{\mu}(t-T)}$ and $\lambda_3 = 1$ with $\sigma = \sqrt{2\mu(2-\mu/L)}$, the constraints of the dual problem (31) can be simplified as

$$e^{-\sigma T} - \eta_1 = 0,$$

$$\begin{bmatrix} 1+\frac{\sigma}{2(L-\mu)}\left(\frac{2\mu}{\sigma}+\sigma\right) & \frac{L\sigma}{L-\mu} & 0 \\ \frac{L\sigma}{L-\mu} & \frac{4\mu L}{L-\mu} & \frac{2\mu}{L-\mu} \\ 0 & \frac{2\mu}{L-\mu} & \frac{2}{L-\mu} \end{bmatrix} - \begin{bmatrix} \frac{L}{L-\mu} & \frac{L\sigma}{L-\mu} & 0 \\ \frac{L\sigma}{L-\mu} & \frac{L\sigma^2}{L-\mu} & 0 \\ 0 & 0 & 0 \end{bmatrix} = \begin{bmatrix} \frac{\mu}{L-\mu}\left(2-\frac{\mu}{L}\right) & 0 & 0 \\ 0 & \frac{2\mu^2}{L-\mu} & -\frac{2\mu}{L-\mu} \\ 0 & -\frac{2\mu}{L-\mu} & \frac{2}{L-\mu} \end{bmatrix} \succeq O,$$

$$\frac{\mu}{2}e^{-\sigma T} - \eta_2 \leq 0,$$

$$\eta_1 \geq 0, \quad \eta_2 \geq 0.$$

Since the second constraint is satisfied, $\eta_1 = e^{-\sigma T}$ and $\eta_2 = \frac{\mu}{2}e^{-\sigma T}$ is a feasible solution. Therefore, we obtain

$$\frac{L}{4(L-\mu)}\left\|\dot{\xi}(T)+\sigma\xi(T)\right\|^2 \leq e^{-\sigma T}\left(f(x(0)) - f^\star + \frac{\mu}{2}\|x(0)-x^\star\|^2\right).$$

■

C cPEP* for gradient norms

In this section, we prove Theorems 3.4 and 3.7. Since the proofs of these theorems involve similar arguments, we first show the common part under a generalized objective function $\mathcal{P}$. A relaxed problem is derived in Appendix C.1 and the Lagrange function is simplified in the first half of Appendix C.2. Then, in Appendices C.2.1 and C.2.2, we show the case specific discussions. Specifically, Appendix C.2.1 shows Theorem 3.7 and Appendix C.2.2 shows Theorem 3.4.

To derive the convergence rate with respect to the gradient norm, we consider the following continuous-time PEP for the ODE (1) with $\beta(t) = 0$:

$$\begin{aligned} \underset{\substack{f\in\mathcal{F}_{\mu,L},\\ x\in C^2([0,T];\mathbb{R}^d)}}{\text{maximize}} \quad & \frac{1}{2}\|\dot{x}(T)\|^2, \\ \text{subject to} \quad & \ddot{x}(t)+\alpha(t)\dot{x}(t)+2\nabla f(x(t)) = 0, \\ & f(x(0)) - f(x(T)) \leq R_1, \\ & \|x(0)-x(T)\|^2 \leq R_2, \quad \dot{x}(0) = 0. \end{aligned}$$

By introducing $\xi(t) \coloneqq x(t) - x(T)$, we rewrite the problem as

$$\begin{aligned} \underset{\substack{f\in\mathcal{F}_{\mu,L},\\ \xi\in C^2([0,T];\mathbb{R}^d),\\ \phi\in C^1([0,T];\mathbb{R}),\\ \gamma\in C([0,T];\mathbb{R}^d)}}{\text{maximize}} \quad & \frac{1}{2}\left\|\dot{\xi}(T)\right\|^2, \\ \text{subject to} \quad & \phi(t) = f(\xi+x(T)) - f(x(T)), \\ & \gamma(t) = \nabla f(\xi+x(T)), \\ & \ddot{\xi}(t)+\alpha(t)\dot{\xi}(t)+2\nabla f(\xi(t)+x(T)) = 0, \\ & \phi(0) \leq R_1, \\ & \|\xi(0)\|^2 \leq R_2, \quad \dot{\xi}(0) = 0. \end{aligned}$$

Note that $\xi(T) = 0$ and $\phi(T) = 0$ hold by the definition of ξ and ϕ, respectively.

C.1 Relaxing the problem

The relaxation below corresponds to that in Section 2.1. This part corresponds to (i) in Section 2.4. However, to derive a good convergence rate on the gradient norm, we consider a different relaxation. Specifically, we can derive a good enough convergence rate without the constraint corresponding to (29) in Appendix B.

From the chain rule, we have

$$\dot{\phi} + \frac{\alpha}{2}\left\|\dot{\xi}\right\|^2 + \frac{1}{2}\left\langle\ddot{\xi},\dot{\xi}\right\rangle = 0, \tag{34}$$

which can be derived in a manner similar to the derivation of (26).

By substituting $u = x(t)$ and $v = x(T)$ to the inequality (27) (this choice corresponds to the case $(\phi', \gamma', \xi') = (0, 0, 0)$, $(\phi'', \gamma'', \xi'') \in \mathcal{G}(T)$ in Section 2.1) and considering the case $L = \infty$, we obtain

$$-\phi(t) - \langle \gamma(t), x(T) - x(t) \rangle \geq \frac{\mu}{2}\mu \|x(T) - x(t)\|^2,$$

which implies

$$\phi(t) - \langle \gamma(t), \xi(t) \rangle + \frac{\mu}{2}\|\xi(t)\|^2 \leq 0.$$

By using the ODE (1), we have

$$\phi + \frac{1}{2}\left\langle \ddot{\xi}, \xi \right\rangle + \frac{\alpha}{2}\left\langle \dot{\xi}, \xi \right\rangle + \frac{\mu}{2}\|\xi\|^2 \leq 0. \tag{35}$$

By using the equality (34) and the inequality (35), we can relax the problem as

$$\begin{aligned}
\underset{\substack{\xi \in C^2([0,T];\mathbb{R}^d),\\ \phi \in C^1([0,T];\mathbb{R})}}{\text{maximize}} \quad & \frac{1}{2}\left\|\dot{\xi}(T)\right\|^2, \\
\text{subject to} \quad & \dot{\phi} + \frac{\alpha}{2}\left\|\dot{\xi}\right\|^2 + \frac{1}{2}\left\langle \ddot{\xi}, \dot{\xi} \right\rangle = 0, \\
& \phi + \frac{1}{2}\left\langle \ddot{\xi}, \xi \right\rangle + \frac{\alpha}{2}\left\langle \dot{\xi}, \xi \right\rangle + \frac{\mu}{2}\|\xi\|^2 \leq 0, \\
& \phi(0) \leq R_1, \\
& \|\xi(0)\|^2 \leq R_2, \quad \dot{\xi}(0) = 0, \quad \xi(T) = 0, \quad \phi(T) = 0.
\end{aligned}$$

C.2 Lagrange dual problem

We define the Lagrange function as

$$\begin{aligned}
\mathcal{L}(\xi, \phi; \lambda_1, \lambda_2, \eta_1, \eta_2) \coloneqq \frac{1}{2}\left\|\dot{\xi}(T)\right\|^2 - \int_0^T \lambda_1(t)\,[\text{LHS of (34)}]\,\mathrm{d}t - \int_0^T \lambda_2(t)\,[\text{LHS of (35)}]\,\mathrm{d}t \\
- \eta_1(\phi(0) - R_1) - \eta_2(\|\xi(0)\|^2 - R_2),
\end{aligned}$$

where $\lambda_1, \lambda_2 \in C^\infty([0,T];\mathbb{R})$ and $\eta_1, \eta_2 \in \mathbb{R}$. The Lagrange dual problem reads

$$\begin{aligned}
\underset{\lambda_1, \lambda_2, \eta_1, \eta_2}{\text{minimize}} \quad & \underset{\substack{\xi \text{ s.t. } \dot{\xi}(0) = \xi(T) = 0\\ \phi \text{ s.t. } \phi(T) = 0}}{\text{maximize}} \quad \mathcal{L}(\xi, \phi; \lambda_1, \lambda_2, \eta_1, \eta_2), \\
\text{subject to} \quad & \lambda_2(t) \geq 0, \lambda_3 \geq 0, \eta_1 \geq 0, \eta_2 \geq 0.
\end{aligned}$$

To simplify the dual problem, we try to eliminate the primal variables x, ϕ.

First, we consider the term $\int_0^T \lambda_1(t)\,[\text{LHS of (34)}]\,\mathrm{d}t$. Using the integration-by-parts formula, $\dot{\xi}(0) = 0$ and $\phi(T) = 0$, we have

$$\begin{aligned}
\int_0^T \lambda_1(t)\,[\text{LHS of (34)}]\,\mathrm{d}t &= \int_0^T \lambda_1 \left(\dot{\phi} + \frac{\alpha}{2}\left\|\dot{\xi}\right\|^2 + \frac{1}{2}\left\langle \ddot{\xi}, \dot{\xi} \right\rangle \right) \mathrm{d}t \\
&= [\lambda_1(t)\phi(t)]_0^T - \int_0^T \dot{\lambda}_1 \phi \,\mathrm{d}t + \frac{1}{2}\int_0^T \alpha\lambda_1 \left\|\dot{\xi}\right\|^2 \mathrm{d}t + \frac{1}{4}\int_0^T \lambda_1 \partial_t \left\|\dot{\xi}\right\|^2 \mathrm{d}t \\
&= -\lambda_1(0)\phi(0) - \int_0^T \dot{\lambda}_1 \phi \,\mathrm{d}t + \frac{1}{2}\int_0^T \alpha\lambda_1 \left\|\dot{\xi}\right\|^2 \mathrm{d}t \\
&\quad + \frac{1}{4}\left[\lambda_1(t)\left\|\dot{\xi}(t)\right\|^2 \right]_0^T - \frac{1}{4}\int_0^T \dot{\lambda}_1 \left\|\dot{\xi}\right\|^2 \mathrm{d}t \\
&= -\lambda_1(0)\phi(0) + \frac{1}{4}\lambda_1(T)\left\|\dot{\xi}(T)\right\|^2
\end{aligned}$$

$$- \int_0^T \dot{\lambda}_1 \phi \,\mathrm{d}t + \frac{1}{4}\int_0^T \Big(2\alpha\lambda_1 - \dot{\lambda}_1\Big)\Big\|\dot{\xi}\Big\|^2 \mathrm{d}t.$$

Second, we consider the term $\int_0^T \lambda_2(t)\,[\text{LHS of (35)}]\,\mathrm{d}t$. Using the integration-by-parts formula, $\dot{\xi}(0) = 0$ and $\xi(T) = 0$, we have

$$\begin{aligned}
\int_0^T \lambda_2(t)\,[\text{LHS of (35)}]\,\mathrm{d}t &= \int_0^T \lambda_2\bigg(\phi + \frac{1}{2}\Big\langle \ddot{\xi}, \xi\Big\rangle + \frac{\alpha}{2}\Big\langle \dot{\xi}, \xi\Big\rangle + \frac{\mu}{2}\|\xi\|^2\bigg)\mathrm{d}t \\
&= \int_0^T \lambda_2\phi\,\mathrm{d}t + \frac{1}{2}\int_0^T \bigg(\partial_t\Big(\lambda_2\Big\langle \dot{\xi}, \xi\Big\rangle\Big) - \dot{\lambda}_2\Big\langle \dot{\xi}, \xi\Big\rangle - \lambda_2\Big\|\dot{\xi}\Big\|^2\bigg)\mathrm{d}t \\
&\quad + \frac{1}{4}\int_0^T \alpha\lambda_2\partial_t\|\xi\|^2\mathrm{d}t + \frac{\mu}{2}\int_0^T \lambda_2\|\xi\|^2\mathrm{d}t \\
&= \int_0^T \lambda_2\phi\,\mathrm{d}t - \frac{1}{2}\int_0^T \lambda_2\Big\|\dot{\xi}\Big\|^2\mathrm{d}t + \frac{\mu}{2}\int_0^T \lambda_2\|\xi\|^2\mathrm{d}t + \frac{1}{4}\int_0^T \Big(\alpha\lambda_2 - \dot{\lambda}_2\Big)\partial_t\|\xi\|^2\mathrm{d}t \\
&= \int_0^T \lambda_2\phi\,\mathrm{d}t - \frac{1}{2}\int_0^T \lambda_2\Big\|\dot{\xi}\Big\|^2\mathrm{d}t + \frac{\mu}{2}\int_0^T \lambda_2\|\xi\|^2\mathrm{d}t \\
&\quad + \frac{1}{4}\Big[\Big(\alpha\lambda_2 - \dot{\lambda}_2\Big)\|\xi\|^2\Big]_0^T - \frac{1}{4}\int_0^T \partial_t\Big(\alpha\lambda_2 - \dot{\lambda}_2\Big)\|\xi\|^2\mathrm{d}t \\
&= -\frac{\alpha(0)\lambda_2(0) - \dot{\lambda}_2(0)}{4}\|\xi(0)\|^2 \\
&\quad + \int_0^T \lambda_2\phi\,\mathrm{d}t - \frac{1}{2}\int_0^T \lambda_2\Big\|\dot{\xi}\Big\|^2\mathrm{d}t + \frac{1}{4}\int_0^T \Big(2\mu\lambda_2 - \partial_t\Big(\alpha\lambda_2 - \dot{\lambda}_2\Big)\Big)\|\xi\|^2\mathrm{d}t.
\end{aligned}$$

Summing up the above two terms, we rewrite the Lagrange function as follows:

$$\begin{aligned}
&\mathcal{L}(\xi, \phi; \lambda_1, \lambda_2, \eta_1, \eta_2) \\
&\quad = \eta_1 R_1 + \eta_2 R_2 + (\lambda_1(0) - \eta_1)\phi(0) - \frac{\lambda_1(T) - 2}{4}\Big\|\dot{\xi}(T)\Big\|^2 + \left(\frac{\alpha(0)\lambda_2(0) - \dot{\lambda}_2(0)}{4} - \eta_2\right)\|\xi(0)\|^2 \\
&\qquad + \int_0^T \Big(\dot{\lambda}_1 - \lambda_2\Big)\phi\,\mathrm{d}t - \frac{1}{4}\int_0^T \Big(2\alpha\lambda_1 - \dot{\lambda}_1 - 2\lambda_2\Big)\Big\|\dot{\xi}\Big\|^2\mathrm{d}t - \frac{1}{4}\int_0^T \Big(2\mu\lambda_2 - \partial_t\Big(\alpha\lambda_2 - \dot{\lambda}_2\Big)\Big)\|\xi\|^2\mathrm{d}t.
\end{aligned}$$

Therefore, the modification (as in Section 2.2) of the Lagrange dual problem reads

$$\begin{aligned}
\underset{\lambda_1, \lambda_2, \eta_1, \eta_2}{\text{minimize}} \quad & \eta_1 R_1 + \eta_2 R_2 \\
\text{subject to} \quad & \lambda_1(0) - \eta_1 = 0, \quad \lambda_1(T) - 2 \geq 0, \\
& \frac{\alpha(0)\lambda_2(0) - \dot{\lambda}_2(0)}{4} - \eta_2 \leq 0, \\
& \dot{\lambda}_1(t) - \lambda_2(t) = 0, \\
& 2\alpha(t)\lambda_1(t) - \dot{\lambda}_1(t) - 2\lambda_2(t) \geq 0, \\
& 2\mu\lambda_2(t) - \partial_t\Big(\alpha(t)\lambda_2(t) - \dot{\lambda}_2(t)\Big) \geq 0, \\
& \lambda_2(t) \geq 0, \quad \eta_1 \geq 0, \quad \eta_2 \geq 0.
\end{aligned}$$

By eliminating λ_2 and η_1 by using $\lambda_2(t) = \dot{\lambda}_1(t)$ and $\eta_1 = \lambda_1(0)$, we obtain

$$\begin{aligned}
\underset{\lambda_1, \eta_2}{\text{minimize}} \quad & \lambda_1(0) R_1 + \eta_2 R_2 \\
\text{subject to} \quad & \lambda_1(T) - 2 \geq 0, \\
& \frac{\alpha(0)\dot{\lambda}_1(0) - \ddot{\lambda}_1(0)}{4} - \eta_2 \leq 0,
\end{aligned}$$

$$
\begin{aligned}
&2\alpha(t)\lambda_1(t) - 3\dot{\lambda}_1(t) \geq 0,\\
&\partial_t\Big(2\mu\lambda_1(t) - \alpha(t)\dot{\lambda}_1(t) + \ddot{\lambda}_1(t)\Big) \geq 0,\\
&\dot{\lambda}_1(t) \geq 0, \quad \lambda_1(0) \geq 0, \quad \eta_2 \geq 0.
\end{aligned}
$$

In the following, we consdier the cases $\mu = 0$ and $\mu > 0$ separately.

C.2.1 Case $\mu = 0$

In this section, we consider the case $\mu = 0$, i.e., Theorem 3.7. This part corresponds to (iii) in Section 2.4. By using ansatz $\alpha(t) = \frac{a}{S-t}$ and $\lambda_1(t) = c(S-t)^p$ with $a, c, p, S \in \mathbb{R}$ $(S > T)$, the constraints can be simplified as

$$
\begin{aligned}
&c(S-T)^p \geq 2,\\
&\frac{1-a-p}{4}cpS^{p-2} - \eta_2 \leq 0,\\
&(2a+3p)c(S-t)^{p-1} \geq 0,\\
&-cp(a+p-1)(p-2)(S-t)^{p-3} \geq 0,\\
&-cp(S-t)^{p-1} \geq 0, \quad cS^p \geq 0, \quad \eta_2 \geq 0.
\end{aligned}
$$

Since we assume $S > T \geq t \geq 0$, these constraints can be further simplified as

$$
\begin{aligned}
&c(S-T)^p \geq 2,\\
&\frac{1-a-p}{4}cpS^{p-2} - \eta_2 \leq 0,\\
&2a+3p \geq 0,\\
&p(a+p-1)(p-2) \leq 0,\\
&p \leq 0, \quad c > 0, \quad \eta_2 \geq 0.
\end{aligned}
$$

Taking $2a + 3p \geq 0$ and $p \leq 0$ together, we see that a must be nonnegative for the feasible region to be nonempty. Then, the constraints $p(a+p-1)(p-2) \leq 0$ and $p \leq 0$ imply that $p \in (-\infty, \min\{0, 1-a\})$. Again, taking $2a + 3p \geq 0$ and $p \in (-\infty, \min\{0, 1-a\})$ together, we see that $a \in [0, 3]$ holds for the feasible region to be nonempty. Under this condition, we should choose $c = 2(S-T)^{-p}$ and $\eta_2 = \frac{1-a-p}{4}(S-T)^p S^{p-2}$, and the dual problem can be simplified as

$$
\begin{aligned}
\underset{p}{\text{minimize}} \quad & 2(S-T)^{-p}S^{p-2}\left(S^2R_1 + p\frac{1-a-p}{4}R_2\right)\\
\text{subject to} \quad & p \in \left[-\frac{2}{3}a, \min\{0, 1-a\}\right].
\end{aligned}
$$

Since S^{p-2} is included in the objective function, p should be smaller, so we choose $p = -\frac{2a}{3}$. Then, the value of the objective function is

$$
2(S-T)^{\frac{2a}{3}}S^{-\frac{2a+6}{3}}\left(S^2R_1 + p\frac{3-a}{12}R_2\right).
$$

To minimize the exponent $-\frac{2a+6}{3}$ under the condition $a \in [0, 3]$, we should choose $a = 3$, and the value of the objective function is

$$
2(S-T)^2S^{-2}R_1.
$$

As a consequence, we obtain the worst-case estimate

$$
\left\|\frac{\dot{x}(T)}{S-T}\right\|^2 \leq 4\frac{f(x(0)) - f(x(T))}{S^2} \leq 4\frac{f(x(0)) - f^\star}{S^2}.
$$

C.2.2 Case $\mu > 0$

In this section, we consider the case $\mu = 0$, i.e., Theorem 3.4. This part corresponds to (iii) in Section 2.4. With reference to the conclusion of the previous section, we use ansatz $\alpha(t) = 3\sqrt{\mu}\coth\big(\sqrt{\mu}(S-t)\big)$ and $\lambda_1(t) = c\sinh^{-2}\big(\sqrt{\mu}(S-t)\big)$. Then, some elementary calculations show that $\dot{\lambda}_1(t) = 2\sqrt{\mu}\coth\big(\sqrt{\mu}(S-t)\big)\lambda_1(t)$ and $\ddot{\lambda}_1(t) = 2\mu\big(3\coth^2\big(\sqrt{\mu}(S-t)\big) - 1\big)\lambda_1(t)$. Under this setting, the constraints can be simplified as

$$
\begin{aligned}
&c\sinh^{-2}(\sqrt{\mu}(S-T)) - 2 \geq 0,\\
&\frac{\coth^2\big(\sqrt{\mu}S\big) - 1}{2}\mu\lambda_1(0) - \eta_2 \leq 0,\\
&(6\sqrt{\mu}\coth(\sqrt{\mu}(S-t)) - 6\sqrt{\mu}\coth(\sqrt{\mu}(S-t)))\lambda_1(t) \geq 0,\\
&\partial_t\big(\big(2\mu - 6\mu\coth^2(\sqrt{\mu}(S-t)) + 2\mu\big(3\coth^2(\sqrt{\mu}(S-t)) - 1\big)\big)\lambda_1\big) \geq 0,\\
&2\sqrt{\mu}\coth(\sqrt{\mu}(S-t))\lambda_1(t) \geq 0, \quad \lambda_1(0) \geq 0, \quad \eta_2 \geq 0.
\end{aligned}
$$

Since the third and forth constraints are obviously satisfied, $c = 2\sinh\big(\sqrt{\mu}(S-T)\big)$ gives a feasible solution. Then, the value of the objective function is

$$
\frac{2\sinh^2\big(\sqrt{\mu}(S-T)\big)}{\sinh^2\big(\sqrt{\mu}S\big)}R_1 + \mu\frac{\sinh^2\big(\sqrt{\mu}(S-T)\big)}{\sinh^4\big(\sqrt{\mu}S\big)}R_2.
$$

As a consequence, we obtain the worst-case estimate

$$
\left\|\frac{\dot{x}(T)}{\sinh\big(\sqrt{\mu}(S-T)\big)}\right\|^2 \leq 4\frac{f(x(0)) - f(x(T))}{\sinh^2\big(\sqrt{\mu}S\big)} + 2\mu\frac{\|x(0) - x(T)\|^2}{\sinh^4\big(\sqrt{\mu}S\big)}.
$$

D Utilizing $\beta(t)$ for further optimizing the rate

We can improve Theorem 3.1 by utilizing $\beta(t)$.

Theorem D.1. *Let $f \in \mathcal{F}_{\mu,\infty}$ and let $x : [0,T] \to \mathbb{R}^d$ be the solution of eq. (1) with $\alpha(t) \coloneqq 2\sigma - 2\mu/\sigma$ and $\beta(t) \coloneqq -\sigma/(2\mu) + 2/\sigma$. Then, if $\sigma = 2\sqrt{\mu}$, we have for any $T \geq 0$*

$$
\begin{aligned}
&f(x(T)) - f^\star - \frac{\mu}{2}\|x(T) - x^\star\|^2 + \frac{1}{4}\|2\sqrt{\mu}(x(T) - x^\star) + \dot{x}(T)\|^2\\
&\quad\leq \mathrm{e}^{-2\sqrt{\mu}T}\Big(f(x(0)) - f^\star + \frac{\mu}{2}\|x(0) - x^\star\|^2\Big).
\end{aligned}
$$

If $\sqrt{(2+\sqrt{2})\mu} \leq \sigma < 2\sqrt{\mu}$, we have for any $T \geq 0$

$$
\begin{aligned}
&f(x(T) + \beta\dot{x}(T)) - f^\star + \frac{\mu\sigma^2}{2(32\mu\sigma^2 - 7\sigma^4 - 16\mu^2)}\left\|\sigma(x(T) - x^\star) + \frac{16\mu\sigma^2 - 3\sigma^4 - 8\mu^2}{2\mu\sigma^2}\dot{x}(T)\right\|^2\\
&\quad\leq \frac{16\mu\sigma^2 - 3\sigma^4 - 8\mu^2}{32\mu\sigma^2 - 7\sigma^4 - 16\mu^2}\mathrm{e}^{-\sigma T}\Big(f(x(0)) - f^\star + \frac{\mu}{2}\|x(0) - x^\star\|^2\Big).
\end{aligned}
$$

Remark D.2. The first inequality is the same as Theorem 3.1 since $\beta(t) = 0$ when $\sigma = 2\sqrt{\mu}$. For the second inequality, the constant before the rate is better than Theorem 3.1 since for $\sqrt{(2+\sqrt{2})\mu} \leq \sigma < 2\sqrt{\mu}$, it holds that

$$
\frac{16\mu\sigma^2 - 3\sigma^4 - 8\mu^2}{32\mu\sigma^2 - 7\sigma^4 - 16\mu^2} < \frac{2\mu}{4\mu - \sigma^2},
$$

where RHS is the constant of Theorem 3.1.

Proof. The first inequality is the same as (21). The second inequality is obtained by using the Lagrange dual problem of (30). By using ansatz $\alpha(t) = 2\sigma - \frac{2\mu}{\sigma}$, $\beta(t) = -\frac{\sigma}{2\mu} + \frac{2}{\sigma}$, $\lambda_1(t) = (1+\lambda_3)e^{\sigma(t-T)}$ and $\lambda_2(t) = (1+\lambda_3)\sigma e^{\sigma(t-T)}$ with $\sigma \geq 0$, the constraints of (30) can be simplified as

$$\begin{aligned}
&(1+\lambda_3)e^{-\sigma T} - \eta_1 = 0,\\
&\begin{bmatrix} (8 - \frac{3\sigma^2}{2\mu} - \frac{4\mu}{\sigma^2})(1+\lambda_3) & \sigma(1+\lambda_3) \\ \sigma(1+\lambda_3) & 2\mu(1+2\lambda_3) \end{bmatrix} \succeq O,\\
&\frac{(1+\lambda_3)\mu}{2} e^{-\sigma T} - \eta_2 \leq 0,\\
&\lambda_3 \geq 0, \quad \eta_1 \geq 0, \quad \eta_2 \geq 0.
\end{aligned}$$

The second constraint is equivalent to $(8 - \frac{3\sigma^2}{2\mu} - \frac{4\mu}{\sigma^2})(1+\lambda_3) \geq 0$, $2\mu(2\lambda_3+1) \geq 0$ and $2\mu(8 - \frac{3\sigma^2}{2\mu} - \frac{4\mu}{\sigma^2})(1+\lambda_3)(\lambda_3+1)(2\lambda_3+1) - \sigma^2(\lambda_3+1)^2 = (\lambda_3+1)\Big(\Big(32\mu - 7\sigma - 16\frac{\mu^2}{\sigma^2}\Big)\lambda_3 + 16\mu - 4\sigma^2 - 8\frac{\mu^2}{\sigma^2}\Big) \geq 0$. As a consequence, when $\sigma \in [\sqrt{(2-\sqrt{2})\mu}, 2\sqrt{\mu})$, $\lambda_1(t) = \frac{16\mu\sigma^2 - 3\sigma^4 - 8\mu^2}{32\mu\sigma^2 - 7\sigma^4 - 16\mu^2} e^{\sigma(t-T)}$, $\lambda_2(t) = \frac{16\mu\sigma^2 - 3\sigma^4 - 8\mu^2}{32\mu\sigma^2 - 7\sigma^4 - 16\mu^2}\sigma e^{\sigma(t-T)}$, $\lambda_3 = \frac{-16\mu\sigma^2 + 4\sigma^4 + 8\mu^2}{32\mu\sigma^2 - 7\sigma^4 - 16\mu^2}$, $\eta_1 = \frac{16\mu\sigma^2 - 3\sigma^4 - 8\mu^2}{32\mu\sigma^2 - 7\sigma^4 - 16\mu^2} e^{-\sigma T}$ and $\eta_2 = \frac{\mu}{2}\frac{16\mu\sigma^2 - 3\sigma^4 - 8\mu^2}{32\mu\sigma^2 - 7\sigma^4 - 16\mu^2} e^{-\sigma T}$ is feasible solution of the dual problem. Therefore, we obtain

$$f(x(T)) - f^\star \leq \frac{16\mu\sigma^2 - 3\sigma^4 - 8\mu^2}{32\mu\sigma^2 - 7\sigma^4 - 16\mu^2} e^{-\sigma T}\Big(f(x(0)) - f^\star + \frac{\mu}{2}\|x(0) - x^\star\|^2\Big).$$

Furthermore, since

$$\begin{bmatrix} (8 - \frac{3\sigma^2}{2\mu} - \frac{4\mu}{\sigma^2})(1+\lambda_3) & \sigma(1+\lambda_3) \\ \sigma(1+\lambda_3) & 2\mu(1+2\lambda_3) \end{bmatrix} = \frac{1}{32\mu\sigma^2 - 7\sigma^4 - 16\mu^2} \begin{bmatrix} \frac{1}{2\mu\sigma^2}(16\mu\sigma^2 - 3\sigma^4 - 8\mu^2)^2 & (16\mu\sigma^2 - 3\sigma^4 - 8\mu^2)\sigma \\ (16\mu\sigma^2 - 3\sigma^4 - 8\mu^2)\sigma & 2\mu\sigma^4 \end{bmatrix}$$

holds for the chosen feasible solution, the same argument holds true even if we replace the objective function with

$$\phi(T) + \frac{\mu\sigma^2}{2(-7\sigma^4 + 32\mu\sigma^2 - 16\mu^2)} \left\| \frac{-3\sigma^4 + 16\mu\sigma^2 - 8\mu^2}{2\mu\sigma^2}\dot{\xi}(T) + \sigma\xi(T) \right\|^2.$$

This implies

$$\begin{aligned}
&f(x(T) + \beta\dot{x}(T)) - f^\star + \frac{\mu\sigma^2}{2(-7\sigma^4 + 32\mu\sigma^2 - 16\mu^2)} \left\| \sigma(x(T) - x^\star) + \frac{-3\sigma^4 + 16\mu\sigma^2 - 8\mu^2}{2\mu\sigma^2}\dot{x}(T) \right\|^2 \\
&\leq \frac{-3\sigma^4 + 16\mu\sigma^2 - 8\mu^2}{-7\sigma^4 + 32\mu\sigma^2 - 16\mu^2} \mathrm{e}^{-\sigma T}\Big(f(x(0)) - f^\star + \frac{\mu}{2}\|x(0) - x^\star\|^2\Big).
\end{aligned}$$

■

E Further results on high-resolution ODEs

In (Shi et al., 2022), the high-resolution ODEs for the heavy-ball method and AGM-SC are derived as follows: for μ-strongly convex f and step size $h > 0$,

- The high-resolution ODE of the heavy-ball method:

$$\ddot{x} + 2\sqrt{\mu}\dot{x} + (1 + \sqrt{\mu}h)\nabla f(x) = 0. \tag{36}$$

- The high-resolution ODE for AGM-SC:

$$\ddot{x} + 2\sqrt{\mu}\dot{x} + h\nabla^2 f(x)\dot{x} + (1 + \sqrt{\mu}h)\nabla f(x) = 0. \tag{37}$$

They showed the following convergence rates.

Proposition E.1 ((Shi et al., 2022)). *Let $f \in \mathcal{F}_{\mu,L}$. For any step size $0 < h \leq 1/\sqrt{L}$, we consider the solutions x of the high-resolution ODEs (36) and (37) with the initial condition $x(0) = x_0$ and $\dot{x}(0) = -\frac{2h\nabla f(x_0)}{1+\sqrt{\mu}h}$. For (36), we have*

$$f(x(T)) - f^\star \leq \frac{7\|x_0 - x^\star\|^2}{2h^2}\mathrm{e}^{-\frac{\sqrt{\mu}}{4}T},$$

and for (37), we have

$$f(x(T)) - f^\star \leq \frac{2\|x_0 - x^\star\|^2}{h^2}\mathrm{e}^{-\frac{\sqrt{\mu}}{4}T}.$$

This theorem implies that the difference between heavy-ball method and AGM-SC appears in the constant factor of the rate, not in the exponent.

By using cPEP*, we can show another convergence rate. In the following theorem, the exponents of the rates differ between the two method. Instead of the initial condition in (Shi et al., 2022), we consider $x(0) = x_0$ and $\dot{x}(0) = 0$ to use the Lagrange function $\mathcal{L}$ calculated in Appendix B.2 where we use $\dot{x}(0) = 0$. We can handle the original initial condition by leaving the term of $\dot{x}(0)$ in the Lagrange function $\mathcal{L}$ as it is.

Theorem E.2. *Let $f \in \mathcal{F}_{\mu,\infty}$. We consider the solutions x of the high-resolution ODEs (36) and (37) with the initial condition $x(0) = x_0$ and $\dot{x}(0) = 0$. For (36), we have*

$$f(x(T)) - f^\star \leq \frac{8}{1+9\sqrt{\mu}h}\left(\frac{9}{8}(1+\sqrt{\mu}h)(f(x(0)) - f^\star) + \frac{\mu}{2}\|x(0) - x^\star\|^2\right)e^{-\frac{4}{3}\sqrt{\mu}T} \quad \text{for any} \quad 0 \leq h, \tag{38}$$

and for (37), we have

$$f(x(T)) - f^\star \leq \left(f(x(0)) - f^\star + \frac{\mu}{2}\|x(0) - x^\star\|^2\right)e^{-(\sqrt{\mu}h+1)\sqrt{\mu}T} \quad \text{for any} \quad 0 \leq h \leq \frac{1}{2\sqrt{\mu}}. \tag{39}$$

Remark E.3. When $h = 1/(2\sqrt{\mu})$, the rate (39) for (37) is the fastest, which is $\mathrm{O}\left(\mathrm{e}^{-\frac{3}{2}\sqrt{\mu}T}\right)$. It might seem that the rate (39) is strange since it suggests $\mathrm{O}\left(\mathrm{e}^{-\sqrt{\mu}T}\right)$ as $h \to 0$, which is weaker than the rate $\mathrm{O}\left(\mathrm{e}^{-\frac{4}{3}\sqrt{\mu}T}\right)$ (38) for (36). This is contradictory, since AGM-SC is faster than the heavy-ball method. This can be resolved if necessary. For (37), we can show another evaluation

$$f(x) - f^\star = \mathrm{O}\left(\mathrm{e}^{-\frac{l^2-l-3+\sqrt{l^4-2l^3+11l^2+22l+9}}{2l}\sqrt{\mu}T}\right),$$

where $l \coloneqq \sqrt{\mu}h$ for $0 \leq h \leq 1/(2\sqrt{\mu})$ (we omit its proof.) This rate coincides with $\mathrm{O}\left(\mathrm{e}^{-\frac{4}{3}\sqrt{\mu}T}\right)$ when $h = 0$ and $\mathrm{O}\left(\mathrm{e}^{-\frac{3}{2}\sqrt{\mu}T}\right)$ when $h = 1/(2\sqrt{\mu})$, and thus it is more consistent with discrete rates. Since the best rates (which are important) for the both evaluation are the same, in the above theorem we avoid the complicated version and show the concise one.

Proof. Let $a \coloneqq \frac{\sqrt{2}}{\sqrt{1+\sqrt{\mu}h}}$. To use the Lagrange function $\mathcal{L}$ in Appendix B.2, we rescale the time by a times so that the coefficient of $\nabla f(x)$ becomes 2. Then, the ODE (36) and (37) reads

$$\ddot{x} + \frac{2\sqrt{2\mu}}{\sqrt{1+\sqrt{\mu}h}}\dot{x} + 2\nabla f(x) = 0,$$

and

$$\ddot{x} + \frac{2\sqrt{2\mu}}{\sqrt{1+\sqrt{\mu}h}}\dot{x} + \frac{\sqrt{2}h}{\sqrt{1+\sqrt{\mu}h}}\nabla^2 f(x)\dot{x} + 2\nabla f(x) = 0.$$

First, we consider the ODE (36). We can use the dual problem derived in Appendix B.2.4: by setting $\alpha(t) = 2a\sqrt{\mu}$ and $\beta(t) = 0$, the problem (30) reads

$$\underset{\lambda_1,\lambda_2,\lambda_3,\eta_1,\eta_2}{\text{minimize}} \quad \eta_1 R_1 + \eta_2 R_2$$

$$
\begin{aligned}
\text{subject to}\quad & \lambda_3 - \lambda_1(T) + 1 = 0, \quad \lambda_1(0) - \eta_1 = 0,\\
& \begin{bmatrix} \lambda_1(T) & \lambda_2(T) \\ \lambda_2(T) & 2a\sqrt{\mu}\lambda_2(T) - \dot{\lambda}_2(T) + 2\lambda_3\mu \end{bmatrix} \succeq O,\\
& \frac{2a\sqrt{\mu}\lambda_2(0) - \dot{\lambda}_2(0)}{4} - \eta_2 \leq 0,\\
& \dot{\lambda}_1(t) - \lambda_2(t) = 0, \quad 2\mu\lambda_2(t) - 2a\sqrt{\mu}\dot{\lambda}_2(t) + \ddot{\lambda}_2(t) \geq 0,\\
& 2a\sqrt{\mu}\lambda_1 - \frac{1}{2}\dot{\lambda}_1 - \lambda_2 \geq 0,\\
& \lambda_2(t) \geq 0, \quad \lambda_3 \geq 0, \quad \eta_1 \geq 0, \quad \eta_2 \geq 0.
\end{aligned}
$$

Using ansatz $\lambda_1 = (1+\lambda_3)\mathrm{e}^{\sigma\sqrt{\mu}(t-T)}$ and $\lambda_2 = (1+\lambda_3)\sigma\sqrt{\mu}\mathrm{e}^{\sigma\sqrt{\mu}(t-T)}$, the constraints can be simplified as

$$
\begin{aligned}
& (1+\lambda_3)e^{-\sigma T} - \eta_1 = 0,\\
& \begin{bmatrix} 1+\lambda_3 & \sigma\sqrt{\mu}(1+\lambda_3) \\ \sigma\sqrt{\mu}(1+\lambda_3) & \sigma(2a-\sigma)\mu(1+\lambda_3) + 2\lambda_3\mu \end{bmatrix} \succeq O,\\
& \frac{\sigma(2a-\sigma)\mu(1+\lambda_3)}{4}e^{-\sigma\sqrt{\mu}T} - \eta_2 \leq 0,\\
& \big(2 - 2a\sigma + \sigma^2\big)\sigma\mu\sqrt{\mu}(1+\lambda_3)e^{\sigma\sqrt{\mu}(t-T)} \geq 0,\\
& \left(2a - \frac{3}{2}\sigma\right)\sqrt{\mu}(1+\lambda_3)e^{\sigma(t-T)} \geq 0,\\
& \lambda_3 \geq 0, \quad \eta_1 \geq 0, \quad \eta_2 \geq 0.
\end{aligned}
$$

Using ansatz $\sigma = 4a/3$ and $\lambda_3 = 4a^2/(9-4a^2)$ and noting $a = \frac{\sqrt{2}}{\sqrt{1+\sqrt{\mu}h}} < \sqrt{2}$, we can show that the second, forth, fifth constraints are satisfied. Then, the constraints on η_1 and η_2 reads

$$
\eta_1 = \frac{9}{9-4a^2}e^{-\frac{4a}{3}\sqrt{\mu}T}, \quad \eta_2 \geq \frac{4a^2}{9-4a^2}\frac{\mu}{2}e^{-\frac{4a}{3}\sqrt{\mu}T}.
$$

Thus, we obtain

$$
f(x(T)) - f^\star \leq \frac{8}{1+9\sqrt{\mu}h}\left(\frac{9}{8}(1+\sqrt{\mu}h)(f(x(0)) - f^\star) + \frac{\mu}{2}\|x(0) - x^\star\|^2\right)e^{-\frac{4a}{3}\sqrt{\mu}T}.
$$

By rescaling the time by $1/a$ times, we obtain the desired result.

Next, we consider the ODE (37). By introducing y such that $x = y + \frac{\sqrt{2}h}{2\sqrt{1+\sqrt{\mu}h}}\dot{y}$, ODE (37) can be rewritten as

$$
\ddot{y} + \frac{2\sqrt{2\mu}}{\sqrt{1+\sqrt{\mu}h}}\dot{y} + 2\nabla f\left(y + \frac{\sqrt{2}h}{2\sqrt{1+\sqrt{\mu}h}}\dot{y}\right) = 0.
$$

Let $l \coloneqq \sqrt{\mu}h$. By setting $\alpha(t) = 2a\sqrt{\mu}$ and $\beta(t) = \frac{al}{2\sqrt{\mu}}$, the problem (30) reads

$$
\begin{aligned}
\underset{\lambda_1,\lambda_2,\lambda_3,\eta_1,\eta_2}{\text{minimize}}\quad & \eta_1 R_1 + \eta_2 R_2\\
\text{subject to}\quad & \lambda_3 - \lambda_1(T) + 1 = 0, \quad \lambda_1(0) - \eta_1 = 0,\\
& \begin{bmatrix} \lambda_1(T)\big(1+a^2l\big) + \lambda_2(T)\frac{al}{2\sqrt{\mu}} & \lambda_2(T) \\ \lambda_2(T) & \lambda_2(T)\big(2a\sqrt{\mu} + al\sqrt{\mu}\big) - \dot{\lambda}_2(T) + 2\lambda_3\mu \end{bmatrix} \succeq O,\\
& \frac{\lambda_2(0)\big(2a\sqrt{\mu} + al\sqrt{\mu}\big) - \dot{\lambda}_2(0)}{4} - \eta_2 \leq 0,\\
& \dot{\lambda}_1(t) - \lambda_2(t) = 0, \quad \lambda_1(t)\frac{al}{2\sqrt{\mu}} \geq 0, \quad 2\mu\lambda_2(t) - \dot{\lambda_2}(t)(2a\sqrt{\mu} + al\sqrt{\mu}) + \ddot{\lambda}_2(t) \geq 0,
\end{aligned}
$$

$$2\lambda_1 a\sqrt{\mu} - \frac{1}{2}\dot{\lambda_1}\big(1+a^2 l\big) + \frac{2\lambda_2\Big(a^2 l + \frac{a^2l^2}{4} - 1\Big) - \dot{\lambda_2}\frac{al}{2\sqrt{\mu}}}{2} \geq 0,$$
$$\lambda_2(t) \geq 0, \quad \lambda_3 \geq 0, \quad \eta_1 \geq 0, \quad \eta_2 \geq 0.$$

Using ansatz $\lambda_1 = (1+\lambda_3)\mathrm{e}^{\sigma\sqrt{\mu}(t-T)}$ and $\lambda_2 = (1+\lambda_3)\sigma\sqrt{\mu}\mathrm{e}^{\sigma\sqrt{\mu}(t-T)}$, the constraints can be simplified as

$$(1+\lambda_3)e^{-\sigma T} - \eta_1 = 0,$$
$$\begin{bmatrix} \big(1+a^2 l + \frac{1}{2}al\sigma\big)(1+\lambda_3) & \sigma(1+\lambda_3) \\ \sigma(1+\lambda_3) & \sigma(2a+al-\sigma)\mu(1+2\lambda_3) + 2\mu\lambda_3 \end{bmatrix} \succeq O,$$
$$\frac{\big(\sigma(2a+al)-\sigma^2\big)\mu(1+\lambda_3)}{4}e^{-\sigma T} - \eta_2 \leq 0,$$
$$\sigma\big(2-(2a+al)\sigma+\sigma^2\big)\mu\sqrt{\mu}(a+\lambda_3)\mathrm{e}^{\sigma\sqrt{\mu(t-T)}} \geq 0$$
$$\left(2a - \frac{1}{2}\left(3 - a^2 l - \frac{1}{2}a^2 l^2\right)\sigma - \frac{1}{4}al\sigma^2\right)\sqrt{\mu}(1+\lambda_3)e^{\sigma(t-T)} \geq 0,$$
$$\lambda_3 \geq 0, \quad \eta_1 \geq 0, \quad \eta_2 \geq 0.$$

Using $a = \frac{\sqrt{2}}{\sqrt{1+\sqrt{\mu}h}}$, the third constraint is equivalent to $\sigma \leq a$ or $a(l+1) \leq \sigma$, and the fourth constraint is equivalent to

$$a\frac{l^2 - l - 3 - \sqrt{l^4 - 2l^3 + 11l^2 + 22l + 9}}{2l} \leq \sigma \leq a\frac{l^2 - l - 3 + \sqrt{l^4 - 2l^3 + 11l^2 + 22l + 9}}{2l}.$$

Using ansatz $\sigma = a(l+1)$, the above inequalies are satisfied if $l \leq \frac{1}{2}$, and the second constraint is satisfied with $\lambda_3 = 0$. Then, the constraints on η_1 and η_2 reads

$$\eta_1 = e^{-a(l+1)\sqrt{\mu}T}, \quad \eta_2 \geq \frac{\mu}{2}e^{-a(l+1)\sqrt{\mu}T}.$$

Thus, we obtain

$$f(x) - f^\star = f\left(y + \frac{\sqrt{2}h}{2\sqrt{1+\sqrt{\mu}h}}\dot{y}\right) - f^\star \leq \Big(f(x(0)) - f^\star + \frac{\mu}{2}\|x(0) - x^\star\|^2\Big)e^{-a(l+1)\sqrt{\mu}T}.$$

By rescaling the time by $1/a$ times, we obtain the desired result.

■

F Newly found Lyapunov functions via cPEP*

Theorem 3.1 and Theorem 3.3 not only give the rates but also Luapunov functions, which reveals the same rate by the Lyapunov approach. Below we demonstrate it.

F.1 New Lyapunov function for TMM ODE

Let us define

$$E := f(x(T)) - f^\star - \frac{\mu}{2}\|x(T) - x^\star\|^2 \quad + \quad \frac{1}{4}\|2\sqrt{\mu}(x(T) - x^\star) + \dot{x}(T)\|^2,$$

which is the LHS of the inequality (21). Below we omit (T) for brevity. We split E into two parts as separated by the long blank spaces above. The first part will appear in the next section again. By differentiating it with respect to T, we have

$$\langle \nabla f(x), \dot{x}\rangle - \mu\langle \dot{x}, x - x^\star\rangle.$$

The second part gives

$$\frac{1}{2}\langle 2\sqrt{\mu}(x - x^\star) + \dot{x}, 2\sqrt{\mu}\dot{x} + \ddot{x}\rangle = \frac{1}{2}\langle 2\sqrt{\mu}(x - x^\star) + \dot{x}, -2\nabla f(x) - \sqrt{\mu}\ddot{x}\rangle$$

$$= -\mu\langle x - x^\star, \dot{x}\rangle - 2\sqrt{\mu}\langle x - x^\star, \nabla f(x)\rangle - \frac{\sqrt{\mu}}{2}\|\dot{x}\|^2 - \langle \dot{x}, \nabla f(x)\rangle,$$

where we used the TMM ODE (5): $3\sqrt{\mu}\dot{x} + \ddot{x} = -2\nabla f(x)$.

Collecting them, we have

$$\begin{aligned}
\dot{E} &= -2\mu\langle \dot{x}, x - x^\star\rangle - \frac{\sqrt{\mu}}{2}\|\dot{x}\|^2 - 2\sqrt{\mu}\langle x - x^\star, \nabla f(x)\rangle \\
&\leq -2\mu\langle \dot{x}, x - x^\star\rangle - \frac{\sqrt{\mu}}{2}\|\dot{x}\|^2 - 2\sqrt{\mu}\Big(f(x) - f^\star + \frac{\mu}{2}\|x - x^\star\|^2\Big) \\
&= -2\sqrt{\mu}\bigg(\sqrt{\mu}\langle \dot{x}, x - x^\star\rangle + \frac{1}{4}\|\dot{x}\|^2 + f(x) - f^\star + \frac{\mu}{2}\|x - x^\star\|^2\bigg), \qquad (40)
\end{aligned}$$

where in the inequality we used the strong convexity of f.

Since

$$\frac{1}{4}\|2\sqrt{\mu}(x(T) - x^\star) + \dot{x}(T)\|^2 = \sqrt{\mu}\langle \dot{x}, x - x^\star\rangle + \frac{1}{4}\|\dot{x}\|^2 + \mu\|x - x^\star\|^2,$$

we have

$$\sqrt{\mu}\langle \dot{x}, x - x^\star\rangle + \frac{1}{4}\|\dot{x}\|^2 + f(x) - f^\star + \frac{\mu}{2}\|x - x^\star\|^2 = f(x) - f^\star - \frac{\mu}{2}\|x - x^\star\|^2 + \frac{1}{4}\|2\sqrt{\mu}(x(T) - x^\star) + \dot{x}(T)\|^2 = E.$$

Thus (40) becomes

$$\dot{E} \leq -2\sqrt{\mu}E,$$

which implies

$$E(t) \leq \mathrm{e}^{-2\sqrt{\mu}t}E(0).$$

This concludes the claim in Theorem 3.1.

F.2 New Lyapunov function for ITEM ODE

Next let us consider Theorem 3.3 and the associated ITEM ODE (7).

Let us define

$$E := f(x) - f^\star - \frac{\mu}{2}\|x - x^\star\|^2 \quad + \quad \frac{1}{4}\|2\sqrt{\mu}\coth(\sqrt{\mu}t)(x - x^\star) + \dot{x}\|^2.$$

It suffices to consider the differentiation of the second part. It gives

$$\begin{aligned}
&\frac{1}{2}\left\langle 2\sqrt{\mu}\coth(\sqrt{\mu}t)(x - x^\star) + \dot{x}, -\frac{2\mu}{\sinh^2(\sqrt{\mu}t)}(x - x^\star) + 2\sqrt{\mu}\coth(\sqrt{\mu}t)\dot{x} + \ddot{x}\right\rangle \\
=&\frac{1}{2}\left\langle 2\sqrt{\mu}\coth(\sqrt{\mu}t)(x - x^\star) + \dot{x}, -\frac{2\mu}{\sinh^2(\sqrt{\mu}t)}(x - x^\star) - 2\nabla f - \sqrt{\mu}\coth(\sqrt{\mu}t)\dot{x}\right\rangle \\
=&-2\mu\sqrt{\mu}\frac{\coth(\sqrt{\mu}t)}{\sinh^2(\sqrt{\mu}t)}\|x - x^\star\|^2 - 2\sqrt{\mu}\coth(\sqrt{\mu}t)\langle x - x^\star, \nabla f\rangle - \mu\coth^2(\sqrt{\mu}t)\langle x - x^\star, \dot{x}\rangle \\
&-\frac{\mu}{\sinh^2(\sqrt{\mu}t)}\langle \dot{x}, x - x^\star\rangle - \langle \dot{x}, \nabla f\rangle - \frac{\sqrt{\mu}}{2}\coth(\sqrt{\mu}t)\|\dot{x}\|^2,
\end{aligned}$$

where in the first equality the ITEM ODE (7): $3\sqrt{\mu}\coth(\sqrt{\mu}t)\dot{x} + \ddot{x} = -2\nabla f$ is used.

Accordingly we have

$$\begin{aligned}
\dot{E} =& -2\mu\sqrt{\mu}\frac{\coth(\sqrt{\mu}t)}{\sinh^2(\sqrt{\mu}t)}\|x - x^\star\|^2 - 2\sqrt{\mu}\coth(\sqrt{\mu}t)\langle x - x^\star, \nabla f\rangle \\
&-\left(\mu\coth^2(\sqrt{\mu}t) + \frac{\mu}{\sinh^2(\sqrt{\mu}t)} + \mu\right)\langle x - x^\star, \dot{x}\rangle - \frac{\sqrt{\mu}}{2}\coth(\sqrt{\mu}t)\|\dot{x}\|^2
\end{aligned}$$

$$= -2\sqrt{\mu}\coth(\sqrt{\mu}t)\left(\frac{\mu}{\sinh^2(\sqrt{\mu}t)}\|x-x^\star\|^2 + \langle x-x^\star, \nabla f\rangle + \sqrt{\mu}\coth(\sqrt{\mu}t)\langle \dot{x}, x-x^\star\rangle + \frac{1}{4}\|\dot{x}\|^2\right). \tag{41}$$

We used the identity $1 + 1/\sinh^2(y) = \coth^2(y)$ that holds for all $y \in \mathbb{R}$.

Since

$$\frac{1}{4}\|2\sqrt{\mu}\coth(\sqrt{\mu}t)(x-x^\star) + \dot{x}\|^2 = \sqrt{\mu}\coth(\sqrt{\mu}t)\langle x-x^\star, \dot{x}\rangle + \mu\coth^2(\sqrt{\mu}t)\|x-x^\star\|^2 + \frac{1}{4}\|\dot{x}\|^2,$$

we have

$$\begin{aligned}
&\frac{\mu}{\sinh^2(\sqrt{\mu}t)}\|x-x^\star\|^2 + \langle x-x^\star, \nabla f\rangle + \sqrt{\mu}\coth(\sqrt{\mu}t)\langle \dot{x}, x-x^\star\rangle + \frac{1}{4}\|\dot{x}\|^2 \\
=&f(x) - f^\star - \frac{\mu}{2}\|x-x^\star\|^2 + \frac{1}{4}\|2\sqrt{\mu}\coth(\sqrt{\mu}t)(x-x^\star) + \dot{x}\|^2 \\
=&E.
\end{aligned}$$

Consequently, (41) becomes

$$\dot{E} \leq -2\sqrt{\mu}\coth(\sqrt{\mu}t)E.$$

Now, we consider $\tilde{E}(t) = \sinh^2(\sqrt{\mu}t)E(t)$ and its derivative.

$$\dot{\tilde{E}}(t) = (2\sqrt{\mu}\sinh(\sqrt{\mu}t)\cosh(\sqrt{\mu}t) - 2\sqrt{\mu}\sinh^2(\sqrt{\mu}t)\coth(\sqrt{\mu}t))E(t) = 0.$$

This implies

$$E(T) = \tilde{E}(T)\sinh^{-2}(\sqrt{\mu}T) = \tilde{E}(0)\sinh^{-2}(\sqrt{\mu}T) = \frac{\mu}{\sinh^2(\sqrt{\mu}T)}\|x(0)-x^\star\|^2,$$

which reproduces the claim of Theorem 3.3.